2008

# An Extended Item Response Theory Model Incorporating Item Response Time

Soo Jeong Ingrisone

FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

# AN EXTENDED ITEM RESPONSE THEORY MODEL INCORPORATING ITEM RESPONSE TIME

By

SOO JEONG INGRISONE

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2008

The members of the Committee approve the Dissertation of Soo Jeong Ingrisone defended on October 23, 2008.

Betsy Jane Becker
Professor Co-Directing Dissertation

Kai-Sheng Song
Professor Co-Directing Dissertation

Fred W. Huffer
Outside Committee Member

Akihito Kamata
Committee Member

Approved:

Akihito Kamata, Chair, Department of Educational Psychology and Learning Systems

The Office of Graduate Studies has verified and approved the above named committee members.

For James and Catherine

# ACKNOWLEDGEMENTS

I would like to express my gratitude to my major professors, Dr. Kai-Sheng Song and Dr. Betsy Becker. Dr. Song has provided many hours of discussions, advice, encouragement, and support that enabled me to find the direction of my research. I am grateful for Dr. Becker's thoughtful comments and her dedication to editing this work which helped me to express my ideas clearly. I am indebted to my outside committee member, Dr. Fred Huffer. His devoted guidance, crucial suggestions and immense patience helped me to advance and complete this dissertation. I am thankful to Dr. Akihito Kamata for introducing me to psychometrics and his insightful comments.

I would like to show my appreciation to my parents for their moral and financial support as well as their love and patience throughout my studies. Special thanks to my husband James and my daughter Catherine. James is my best friend, biggest supporter, most trusted colleague, and best manuscript reviewer that I know. I lack the words to praise the love that he has bestowed upon me. And to my lovely daughter Catherine, I thank her for sharing this moment with me. My life is filled with blessings because of her.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# ABSTRACT

There is a growing need to use response time data to improve measurement quality with the increasing popularity of computerized testing. This work simultaneously models item response and response time to improve on current IRT models that do not account for response time when there is a time limit in real testing. The joint distribution for item response and response time is presented in this work. It is specified as the product of the conditional distribution of response accuracy given response time and the marginal distribution of response time based on the lognormal distribution. A modified version of Thissen's (1983) log linear model is used to fit the response time. Marginal maximum likelihood estimation is developed and employed to estimate the item parameters. In addition, a maximum a posteriori procedure is developed and implemented to estimate person parameters. Three different simulation studies were conducted to evaluate the precision of estimation procedures. The results of item and person parameter estimates based on MML and MAP procedures were found to be consistent and accurate.

# CHAPTER 1

# INTRODUCTION

Response time data has largely been ignored in the field of educational measurement. The main reason is that it has been difficult to collect response time data at the individual item level with paper-based testing. Recently, due to the use of computer technology in testing, in addition to response accuracy data, response time data can be obtained at the item level. As a result, the amount of time examinees spend on each item can be investigated. Ease of measurement and the availability of the data have led to growing interest in the use of response time in the measurement field (Schnipke & Scrams, 2002; Wang, 2006; Wang & Zhang, 2006).

The analysis of response time however has had a long history in cognitive psychology due to the fact that it is a ubiquitous dependent variable in the field (Luce, 1986). In any choice of psychological experiment, there are at least two dependent measures, that is, the choices a subject makes and the time a subject takes. Thus, psychologists have speculated that response time may reveal information about mental processes which require different amounts of time. In other words, how long it takes a person to process a task reveals how the person processed it. For the most part, the distribution of response time has been inferred as a source of information about how the mind processes. In particular, the Speed-Accuracy Tradeoff Function (SATF) describes the compromises a subject makes between accuracy and time demands. SATF analysis indicates how a subject’s accuracy in performing certain tasks changes as the response time changes (Luce, 1986). Therefore, separate analyses of response accuracy and time have been shown to be misleading (Thissen, 1983).

Most research regarding classical test theory and item response theory has been focused exclusively on response accuracy because it is assumed that response time and accuracy are measuring the same constructs, so that time limits will not affect the response accuracy. Yet, the implication is that an examinee's ability may be measured on the scale of accuracy, the scale of speed, or some combination of the two (Schnipke & Scrams, 2002). In the context of educational measurement, this is a serious issue for the validity of tests. Until now, the time limits have mainly been dealt with as an issue regarding administrative convenience in the context of aptitude and achievement testing. However, several studies demonstrate that speed and accuracy of complex tasks, e.g., in aptitude and achievement testing, do not measure same construct. The research on response time in testing has uncovered a rather convoluted relationship between response time and response accuracy among examinees (Schnipke & Scrams, 2002).

Likewise, the application of conventional item response theory (IRT) models may not be appropriate for real testing settings, because item response theory implicitly assumes nonspeededness (Hambleton & Swaminathan, 1985). Item response theory fundamentally assumes that the test is a pure power test (Roskam, 1997). Unlike the pure speed tests which typically contain test items that are fairly easy, so examinees will almost always answer the items correctly with unlimited time, the pure power tests contain items of varying difficulty so that even if the examinee has unlimited testing time, the examinee will not always answer all the items correctly. Based on this assumption, conventional IRT models only the response accuracy, ignoring response time. In practice, however, pure power tests or pure speed tests are rarely employed (Lord & Novick, 1968). Most existing tests are "hybrid" in nature (van der Linden & Hambleton, 1997, p. 166). In common educational testing, most power tests involve a speed component, namely, time limits. Therefore, the assumption of the nonspeededness is likely violated to various degrees in practice (Oshima, 1994).

Up until now studies have revealed that: Time limits have an effect on examinee performance (Hopkins, 1998); different time limits result in different test scores for examinees (Bridgeman et al., 2003; Bridgeman et al., 2004); speededness affects the estimation of ability and item parameters (Oshima, 1994); IRT item parameter estimates are distorted by speededness (Schnipke, 1996); and ignoring response time data will have an adverse affect in estimating examinee ability (Wang & Hanson, 2005). Schnipke and Scrams (2002, p. 247) confirm that "The IRT ability estimate… represents the examinees' accuracy given the time constraints of the administration; this is clearly confounded with response speed. This is a serious confound and a solution is needed." There is a growing need to use response time data to improve measurement quality with the ever-increasing popularity of computer-based testing (Wang & Hanson, 2005). Consequently, developing a realistic model for tests which incorporates response time as well as applies to timed power tests is needed (van der Linden & Hambleton, 1997). By taking response time into account in the calibration process, this research can facilitate the improvement of ability and item parameter estimation.

The primary purposes of this study are as follows: first, to suggest an item response model that incorporates response time that can be applied to timed power tests, second, to suggest the estimation procedures to calibrate item parameters and to estimate person parameters, and last, to use a simulation to evaluate the model and the parameter estimation procedures.

We may be obligated to investigate response times in the interest of fairness and equity (Schnipke & Scrams, 2002). The proposed model incorporating response time is an important step in the field which can be an aid to enhanced measurement quality.

# CHAPTER 2

## REVIEW OF LITERATURE

Most of the previous research regarding response time modeling has treated response time as a dependent variable (Schnipke & Scrams, 2002). Thissen (1983) also points out that previous response time models offered have ignored issues of response accuracy by not considering item responses simultaneously (Tatsuoka & Tatsuoka, 1980) or by dealing with relatively uncomplicated cognitive tasks (Samejima, 1973, 1974, 1983; Schleiblechner, 1979, 1985). Otherwise, response time is not treated as a dependent variable, but used rather as a fixed variable in models for the prediction of item response (White, 1973, 1979). Likewise, item responses are not considered simultaneously. However, there are a few exceptions where response accuracy and response time are modeled simultaneously. These include Thissen (1983), Roskam (1997), Verhelst, Verstralen, and Jansen (1997), Wang and Hanson (2005) and Wang (2006).

Roskam (1997) and Verhelst et al. (1997) have several similarities: First, they both use Rasch models. In addition, both of their promising models assume the Conditional Accuracy Function (CAF) mechanism. The CAF represents the probability of a correct response conditional on the response time within a fixed speed-accuracy tradeoff (SAT). By assuming CAF as an increasing function, their models show that as response time goes to infinity, the probability of correct response approaches one. In other words, if the examinee has unlimited testing time, the examinee will almost always answer the items correctly. Consequently, their models are only true in the context of speed tests. In addition, they assume that CAF is independent of the subjects' strategy. Namely, a strategy parameter is seen as influencing the response time distribution, not the probability of correct response. Thus, the probability of correct response is governed by the strategy of the subject only through the response time distribution.

Roskam (1997) presents a Rasch-Weibull model. In his model, response time is used to predict finishing time of a test. The probability of correct response is mentioned as a Rasch response time model where the correct response probability is conditioned on response time. It is expressed as

$$P\left(U_{ij}=1\middle|t_{ij},j,i\right)=\frac{\theta_j t_{ij}}{\theta_j t_{ij}+\varepsilon_i}=\frac{\exp\left(\xi_j+\tau_{ij}-\sigma_i\right)}{1+\exp\left(\xi_j+\tau_{ij}-\sigma_i\right)}, \tag{2.1}$$

where $\theta_j$ is called as a mental speed, $t_{ij}$ is response time, $\varepsilon_i$ is item difficulty, and the corresponding logarithms are $\xi_j, \tau_{ij}$, and $\sigma_i$ respectively. In the Rasch response time model, $\theta_j t_{ij}$ is the effective ability parameter. The effective ability parameter for item $i$ is a function of mental speed as well as persistence in attempting to solve an item (processing time). In other words, a person's effective ability solving an item increases as the time invested in solving it increases. According to Roskam (1997), the rate of that increase is a mental speed. Therefore, in the effective ability parameter, the task parameter (mental speed) and strategy parameter (Speed-Accuracy Tradeoff) are confounded. Moreover, Roskam (1997) shows that his CAF Rasch response time model coincides with SATF, in that its shape is determined by mental speed and item difficulty, and the actual trade-off between precision and speed is determined by the examinee's persistence.

Roskam (1997) introduces the marginal probability of correct response. The marginal probability of correct response is obtained by multiplying the Rasch response time model by the Weibull density, and then integrating response time out. The probability is

$$P\left(U_{ij}=1\middle|j,i\right)=\int_0^\infty \frac{\theta_j t}{\theta_j t+\varepsilon_i} f\left(t\right)dt. \tag{2.2}$$

He claims that the equation 2.2 is approximately a Rasch model. It implies that the Rasch homogeneity can be present across persons who invest different amounts of time in each item. Thus, his Weibull distribution model involves test completion time for the entire test rather than individual item response time. A Weibull distribution is used to specify the response time distribution. It is given as

$$f(t) = \lambda \exp\left(-\frac{\lambda}{2}t^2\right), \text{ where } \lambda = \frac{\theta_j}{\varepsilon_i \delta_j}. \tag{2.3}$$

The hazard function is defined as

$$h_{ij}(t) = \frac{\theta_j}{\varepsilon_i \delta_j} t. \tag{2.4}$$

The response time distribution is characterized by the hazard function $h_{ij}(t)$ which is a function of mental speed $(\theta_j)$, examinee's persistence $(\delta_j)$ and the item difficulty $(\varepsilon_i)$. The model behaves such that examinees will spend more time on harder the items, mentally faster examinees will spend less time, and more persistent examinees will spend more time. In addition, a more persistent examinee will invest more time for each item and will increase the probability of correct responses.

The limitation of Roskam's (1997) approach is that his model can only be applied to speed tests due to his crucial assumption that the probability of correct response increases to unity as the response time increases to infinity. Also, he builds his model based on the assumption that response time at the item level is not available in the standard time limit tests, so that only the total test time can be considered. His model makes sense when all items have equal difficulty so response time at the item level is not important. In typical educational tests, however, items are at various item difficulty levels. Therefore, this model cannot be applied to many educational tests. In addition, his response time distribution uses only the one parameter Weibull distribution, thus, it is a narrow application of the Weibull distribution. It does not provide a real response time distribution, rather a limited scope of response time distribution possibilities. Moreover, the validity of his model is disputable. Roskam (1997) admits that there are currently no empirical examples available to exhibit the entire characteristics of Rasch-Weibull model. Results establishing the validity of his entire model were not presented in his study.

Verhelst et al. (1997) focuses on a Rasch model for the marginal probability of a correct response, integrating over response time. In the typical IRT models, the item response function for the binary responses $f_i\left(\theta_j\right)$ is the probability of correct response on item $i$ as a function of the latent variable $\theta$. The item difficulty parameter $\varepsilon_i$ corresponds to the value of the latent variable $\theta$ where the probability of correct response is 0.5. Similarly, Verhelst et al. (1997) assume that the mental activities of the examinee result in some value z which is from random variable Z. If the item is answered correctly, the value z is larger than threshold $\varepsilon_i$. He assumes that Z belongs to a shift family, thus the person parameter $\theta$ is defined to be a location parameter. It is expressed as

$$f_i\left(\theta_j\right) = P_{\theta_j}\left(Z > \varepsilon_i\right) = \int_{\varepsilon_i}^{\infty} g_{\theta_j}\left(z\right)dz, \tag{2.5}$$

where $g_\theta\left(\cdot\right)$ is the logistic probability density function (*pdf*) which results in a logistic IRT model. A random variable $Z$ is called the momentary ability. Assuming a random variable $Z$ takes some value $z$, the item will be correctly answered when the examinee's momentary ability exceeds item difficulty (See Equation 2.5). According to Verhelst et al. (1997), the amounts of time spent to answer the item will explain the variation of the momentary ability *Z*. It is stated that examinee's momentary ability depends on examinee's mental power as well as speed. In addition, the momentary ability is defined as an increasing function of time. Thus, as an examinee spends more time on an item, the probability of a correct response increases. The marginal distribution of *Z* with time integrated out is given by

$$g_\theta\left(z\right) = \int_0^{\infty} h_\theta\left(z \middle| t\right) q_\lambda\left(t\right)dt. \tag{2.6}$$

Verhelst et al. (1997) believe that each examinee decides on the amount of time to spend on an item, which can be represented as a two parameter gamma distributed response time. The respective time distribution is given by

$$q_\lambda\left(t\right) = \frac{\beta^p}{\Gamma\left(p\right)} t^{p-1} e^{-\beta t}, \qquad \lambda = \left(\beta, p\right). \tag{2.7}$$

where the scale parameter $(\beta)$ is an examinee parameter and the shape parameter $(p)$ is considered as the item parameter. Verhelst et al. (1997) state that the momentary ability conditioned on response time can be represented by a generalized form of the extreme value distribution. Its distribution function is expressed as

$$\int_0^\infty h_\theta\left(z\middle|t\right)dt = 1 - \exp\left\{-t\alpha\exp\left[\frac{(\theta-\varepsilon_i)}{\alpha}\right]\right\}, \qquad \alpha > 0. \tag{2.8}$$

where $\alpha$ is conceived of as a constant. With gamma distributed response time distribution, the marginal momentary ability distribution is a generalized logistic function. Therefore, this model is consistent with the logistic item response function. When the shape parameter is equal to one, the response time distribution reduces to an exponential distribution and the marginal momentary ability distribution reduces to the Rasch model. Verhelst et al. (1997) label $\theta/\alpha$ as "mental power" which stands for the combination of speed and accuracy.

According to the simulation results presented by Verhelst et al. (1997), the estimation of the fundamental concept mental power is biased and $\theta/\alpha$ is systematically underestimated. They also report that empirical research using their model is a challenging endeavor.

Verhelst et al. (1997) take both correctness and response time simultaneously into account in their model. However, their model can be only applied to speed tests. They assume the speed-accuracy tradeoff mechanism in their proposed model. That is to say, the probability of correct response goes to one as the amount of time spent in answering an item increases without bounds. As noted above, this is only possible in the context of speed tests where items are relatively easy and errors are mainly caused by time pressure. This assumption is unrealistic in power tests. In addition, when Verhelst et al.'s (1997) Rasch model is compared with the Rasch-Weibull model by Roskam (1997), it is found to be a close approximation of the Rasch-Weibull model. However, Roskam (1997) indicates that the parameter $\alpha$ in this model seems to be unidentifiable and lacks a clear interpretation.

Thissen (1983), Wang and Hanson (2005), and Wang (2006) proposed other models that apply to power tests with time limits. Thissen (1983) presented an item response model as a joint distribution in which the marginal distribution of correct responses is characterized in terms of a two parameter logistic model (2 PL) and the marginal distribution of response time is characterized as a lognormal model. His was the first attempt at incorporating response latency in the context of item response theory as a joint distribution in timed testing. His approach is based on the response latency theory developed in cognitive psychology where response latency is a central part of cognitive processing (Luce, 1986). Although the focus of cognitive psychologists has been the cognitive processes within-person relationship between speed and accuracy, Thissen's (1983) model represents the across-person relationship between speed and accuracy which is a psychometric perspective for the speed-accuracy relationship (Scrams & Schnipke, 1997). His proposed model is built on a revised version of Furneaux's (1961) model.

Based on the assumption of independence of correct responses and response times, Thissen (1983) proposed the joint distribution between correct response and response time as the product of the marginal distribution of two variables. The probability of correct responses $r_{ij}$ = 1 for person $i$ to item $j$ is expressed using the logistic model

$$P\left(r_{ij}=1\right)=\frac{1}{1+\exp\left(-z_{ij}\right)}, \text{ where } z_{ij}=a_j\theta_i+c_j. \tag{2.9}$$

Thissen (1983) called $\theta_i$ as the effective ability of person $i$ in order to distinguish it from the conventional ability estimates obtained with speeded tests, $a_j$ is the item discrimination parameter or slope of item $j$, and $c_j$ is the easiness of item $j$.

Thissen (1983) assumes a lognormal distribution for the expected response time distribution for an examinee responding to an item. He believes that the response time of person $i$ on item $j$, $t_{ij}$, must be a function of parameters representing person $i$ and item $j$ characteristics because the response accuracy or latency or both can be contained in the item responses. The linear model is proposed to portray attributes of both examinees and items that contribute to latency but unrelated to trait the items are intended to measure. The log-linear model is expressed as

$$\ln\left(t_{ij}\right) = \upsilon + s_i + u_j - bz_{ij} + \varepsilon_{ij}, \tag{2.10}$$

where the natural logarithm of response time, $\ln\left(t_{ij}\right)$, is described as a linear function of examinee and item parameters. $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$, $\upsilon$ is the overall mean, $s$ and $u$ are person and item slowness parameters, respectively. The parameter $b$ denotes a regression coefficient representing the log-linear relationship between response time and examinee ability. As $z_{ij}$ increases (as person ability and item easiness increase), response latency decreases. Due to this reason, the relationship between effective ability and slowness can be viewed as another facet of the speed-accuracy tradeoff.

Thissen applied this model to a set of data. His examples illustrate the potential problems associated with a timed testing situation. Overall, the goodness of fit test of the model using $\chi^2$ reveals that his timed testing model is correct. One interesting finding is that correct responses take less time than incorrect responses. That is, correct responses are related to shorter latencies. His results also demonstrate that there exist some processing differences between correct and incorrect responses. Thissen (1983) concludes that all analyses of the test scores in timed testing should be two dimensional. In any timed testing environment, a valid response variable absorbs effective ability or slowness or both. This leads to the conclusion that univariate analysis results regarding effective ability or slowness may be misleading. Therefore, the analysis of a two dimensional response space is needed to estimate the ability parameter.

Although Thissen (1983) applied his model to a speed test, it can also be applied to timed large scale, standardized achievement and aptitude tests (Schnipke & Scrams, 2002). He indicates that the purpose of his model is to provide a practical description of responses to test items rather than an explanation of the cognitive processes underlying these responses. He suggests that process models should be developed in future research. Yet the limitation of Thissen's (1983) model is that it assumes correct responses and response times to be independent. His model assumption is problematic because the two marginal distributions share some common parameters, such as ability, item discrimination and item easiness. In addition, the person slowness parameter $s$, the rate of work, can be seen as a personality trait. Thus, response time results from a person parameter (Schnipke & Scrams, 2002). Consequently, accuracy of response and response time are not independent, thus his assumption does not hold in general. van der Linden and van Krimpen-Stoop (2003) point out that another assumption that Thissen (1983) introduced in his response model is inaccurate. Namely, Thissen (1983) assumes a monotonically decreasing relation between slowness and ability by stating that "more able students work faster" (Thissen, 1983, p. 202). However, van der Linden and van Krimpen-Stoop (2003) found that the literature regarding response latency seems to imply a monotonically increasing function. When an examinee selects the option as accuracy rather than speed, both the value of ability and slowness parameter (that appears in 2.10) seem to increase. Also, empirical evidence supports that the relation between ability and slowness depends on the level of speededness of the test (van der Linden & van Krimpen-Stoop, 2003).

Wang and Hanson (2005) offer a four parameter logistic response time (4PLRT) model. This model is specifically formulated for power tests. In particular, as response times goes to infinity, the model reduces to three parameter logistic model (3PL). In that regard, the 4PLRT model can be viewed as an extension of the conventional 3PL model, incorporating response time. In the 4 PLRT model, response time is an independent variable that affects the probability of a correct response, namely, via the conditional distribution of response accuracy given response time.

In the 4 PLRT model, the probability of a correct response to item $j$ by examinee $i$ is given as

$$P\left(x_{ij}=1\middle|\theta_i,\rho_i,a_j,b_j,c_j,d_j,t_{ij}\right)=c_j+\frac{1-c_j}{1+e^{-1.7a_j\left[\theta_i-\left(\rho_i d_j/t_{ij}\right)-b_j\right]}}, \qquad (2.11)$$

where $a_j$, $b_j$ and $c_j$ are item discrimination, difficulty and guessing parameters respectively, and $\theta_i$ is the person ability parameter. These four parameters are the usual 3PL item response theory (IRT) parameters. Also, 2.10 includes $d_j$ as an item slowness parameter and $\rho_i$ as a person slowness parameter, which are unique to the 4PLRT model. These two parameters are called slowness parameters because the larger that $d_j$ and $\rho_i$ are, the slower the probability converges to its asymptote $a\left(\theta - b\right)$. According to Wang and Hanson (2005), the item slowness parameter shows how items behave to response time. It only relates to a particular item and does not vary across examinees. Similarly, the person slowness parameter indicates the pace of an examinee to answer any item correctly. It only relates to a particular examinee and does not change across items. Thus, there is no interaction between an examinee and an item as stated by Wang and Hanson (2005). Wang and Hanson (2005) assume that the product of these two slowness parameters determines the rate of probability change over response time for a particular examinee to a particular item. By putting a negative term containing the inverse of response time in the exponent of the logistic function, they ensure that the exponent does not increase to one but converges to $a\left(\theta - b\right)$. That is, as response time approaches to infinity, correct response probabilities do not converge to one, but to values less than one defined in the 3 PL model.

Since time limits are common in real testing situations, the 4PLRT model provides promising capabilities in using response time data in time constrained power tests. Likewise, it may explain item response behavior more accurately than a 3PL model. Wang and Hanson (2005) conclude that an estimation procedure using the EM algorithm works well and produces reasonably accurate parameter estimates. The results based on real test data demonstrate that the item parameter estimates from 4PLRT are comparable to those from the 3PL model. Moreover, the in results from simulated data exhibit that when the 4PLRT model fits the data, ignoring response time data will undesirably affect the estimation of examinee abilities.

The 4PLRT model proposed by Wang and Hanson (2005), however, contains some inherent problems. Namely, for the sake of parameter estimation, their technical procedure requires an assumption that the item response time is independent of person parameters. Their rationale is that the person slowness parameter entails how response time affects the probability of the correct answer. Thus, response time can be controlled by either examinees or the test giver. This assumption does not hold in general because response time seems like to be connected to person parameters. Wang and Hanson (2005) realize that the model is justified only when the amount of response time spent on each item is not controlled by the examinees, but by the test giver. Due to that reason, the response time is treated as a fixed predictor rather than a random variable. Wang and Hanson (2005) argue that response time is usually treated as a random variable because when the testing process for the same group of examinees and for the same test is repeated, response times for these items change. Therefore, treating the fundamentally random variable response time as a fixed variable would introduce more bias to the parameter estimation. Consequently, although they avoid modeling response time with this assumption, it causes a serious limitation to the applicability of their model.

To overcome this limitation of their model, Wang and Hanson (2005) suggested modeling the joint distribution of response accuracy and response time. They recommend the joint distribution as the product of the conditional distribution of correct response and the marginal distribution of response time. The complete model can treat response time as a random variable so that it can be applied to testing situations where examinees have full control of how much time they would like to spend on each item up to the time limit.

Wang (2006) presents a model for the joint distribution of response accuracy and response time. His approach is an extension of Wang and Hanson's (2005) 4PLRT model. By developing a joint distribution, Wang's (2006) model does not assume that response time is independent of person parameters, like in Wang and Hanson (2005). Thus, his model expands the applicability of the model. Wang (2006) specifies his joint distribution as the product of the conditional distribution of response accuracy and the marginal distribution of response time. The joint distribution of $y_{ij}$ and $t_{ij}$ is specified as

$$f\left(y_{ij},t_{ij}\left|\theta_i,\rho_i,\delta_j\right.\right)=f\left(y_{ij}\left|t_{ij},\theta_i,\delta_j\right.\right)f\left(t_{ij}\left|\theta_i,\rho_i,\delta_j\right.\right), \qquad (2.12)$$

where $y_{ij}$ is the dichotomous item response variable, with $y_{ij}$ = 1 for correct response and 0 for an incorrect response, and $t_{ij}$ is the response time variable. Here, $\theta_i$ and $\rho_i$ are examinee ability and speed parameters respectively, and $\delta_j=\left(a_j,b_j,c_j,d_j\right)$ are item discrimination, difficulty, guessing and slowness parameters for item *j*, respectively. The specified parameters have the same characteristics as described in Wang and Hanson (2005) except $\rho_i$. The conditional distribution of $y_{ij}$ given $t_{ij}$ is given as

$$f\left(y_{ij}\left|t_{ij},\theta_i,\delta_j\right.\right)=P\left(t_{ij},\theta_i,\delta_j\right)^{y_{ij}}\left[1-P\left(t_{ij},\theta_i,\delta_j\right)\right]^{1-y_{ij}}, \qquad (2.13)$$

where the probability of a correct response to item *j* by examinee *i* is expressed by

$$P\left(t_{ij},\theta_i,\delta_j\right)=\mathrm{P}\left(y_{ij}=1\left|t_{ij},\theta_i,\delta_j\right.\right)=c_j+\frac{1-c_j}{1+e^{-1.7a_j\left[\theta_i-\left(d_j/t_{ij}\right)-b_j\right]}}. \qquad (2.14)$$

Wang (2006) notes that in comparison to Wang and Hanson's (2005) model the difference is that $\rho_i$ is omitted in the exponential function in the model.

For the marginal distribution of response time, he uses a one parameter Weibull distribution. It is a special case of the three parameter Weibull distribution and is limited by setting the location parameter to zero and shape parameter to two. That means that he constrains the response time distribution to be a linear function of time. The marginal distribution of response time is expressed as

$$f\left(t_{ij}\left|\theta_i,\rho_i,\delta_j\right.\right)=\lambda t_{ij}e^{-\lambda t_{ij}/2}\text{, where }\lambda=\rho_i\left(\theta_i-b_j\right)^2, \qquad (2.15)$$

where $\lambda$ is the scale parameter that determines the mean and variance of the response time distribution. The underlying assumption of Wang's (2006) $\lambda$ parameter model is consistent with Wang and Zhang's (2006) finding that examinees will spend more time on items that correspond to their ability level. This can be shown because when $\lambda$ becomes large the mean and variance of the response time distribution becomes small. Also, when $\rho_i$ or the difference between $\theta_i$ and $b_j$ are large then $\lambda$ becomes large. Therefore, in Wang's (2006) view, $\rho_i$ is a speed parameter.

Wang (2006) applies his model to the same real test data as Wang and Hanson (2005). His item parameter estimates are somewhat different from the values reported in Wang and Hanson (2005). As Wang (2006) admits, it should be addressed how much of differences are acceptable when different models are applied to the same data. His simulation study showed that the calibration procedure recovered true item parameters pretty well except in a few cases.

The main limitation of Wang's (2006) model is due to his response time model. His one parameter Weibull distribution is an oversimplified version of the response time distribution, thus it does not adequately capture the realistic representation of response time distribution. However, his approach can be viewed as an initial step for towards simultaneous modeling of response time and response accuracy.

Ingrisone (2008) developed a joint distribution which simultaneously models response accuracy and response time. His conditional distribution incorporates response time into one parameter logistic model. He extends Wang's (2006) model and improves it by using a two parameter Weibull distribution for the marginal distribution of response time. By modeling the shape and scale parameters, the two-parameter Weibull distribution presents a more realistic picture of the response time relationships.

It is challenging, but important to develop a new model regarding the response accuracy and response time problem. The goal of the present work is to simultaneously model item response accuracy and response time in the real testing context. To do this, a joint distribution of item response and response time is used. The proposed model in this work introduces an extended IRT model for the probability of correct response conditioned on response time. For response time distribution, the lognormal distribution is assumed. In addition, a modified version of Thissen's (1983) log linear model is employed to fit the response time. In the following methods section, the proposed model is described in more detail.

# CHAPTER 3

# METHODS

This chapter includes a brief overview of typical item response theory for binary data and item response function for the 2 PL model. Then, a new model is introduced, an extended item response model that incorporates response time. In the present work, a joint distribution for item response and response time is suggested. The model is specified as the product of the conditional distribution of response accuracy given response time and the marginal distribution of response time based on the lognormal distribution. In addition, a modified version of Thissen's (1983) log linear model is employed to fit the response time. In addition, the joint likelihood of item responses and response times are derived. In order to estimate item parameters, marginal maximum likelihood approach is presented. Maximum a posteriori procedure is included to estimate person parameters. Finally, three different simulation studies are arranged to evaluate the precision of item and person parameter estimates.

## Item response theory for binary data

Item response theory is a statistical model-based measurement theory to describe the relationship between observed examinee test performance and an underlying examinee ability level, known as $\theta$, often using a logistic function of one or more item parameters for each item (Folk & Smith, 2002; Hambleton & Swaminathan, 1985; Weiss & Yoes, 1990). The mathematical form of this relationship is specified as a monotonically increasing item response function also known as an item characteristic curve (ICC) for the item (Lord, 1980). The frequency distribution of item scores with binary responses (1 if correct; 0 if incorrect) for fixed ability $\theta$ is expressed as

$$f_j\left(u_j|\theta\right) = P_j\left(\theta\right)^{u_j} Q_j\left(\theta\right)^{1-u_j}, \quad \text{thus}$$

$$\begin{aligned} f_j\left(u_j|\theta\right) &= P_j\left(\theta\right) \quad \mathit{if} \quad u_j = 1 \\ f_j\left(u_j|\theta\right) &= Q_j\left(\theta\right) \quad \mathit{if} \quad u_j = 0, \end{aligned} \tag{3.1}$$

where for a fixed ability $\theta$, the value $u_j = 1$ is termed a “correct” response by an examinee to an item and $P_j$ is referred to as the probability of correct response by an examinee to item *j*. The value $u_j = 0$ is termed an “incorrect” response by an examinee to an item and $Q_j = 1 - P_j$ is referred to as the probability of incorrect response by an examinee to item *j*. The curve connecting the means of the above conditional distributions to the fixed ability $\theta$ is the regression of item score on ability (Hambleton & Swaminathan, 1985). It is referred to as item characteristic curve (ICC). Namely, as the ability level increases, the probability of a correct response to an item increases. The core of the IRT function expresses the probability of observing a particular item response, given an examinee’s ability value and the item parameters (Yen & Fitzpatrick, 2006). For example, examinees with high ability levels will have higher expected probabilities of answering an item correctly than do those with low ability levels.

IRT models include a set of stringent assumptions about the binary data to which the model is applied (Hambleton et al., 1991; Weiss & Yoes, 1990). Item response models that assume a single dominant latent ability which is measured by the items and explains the performance on the test are referred to unidimensional. Another assumption equivalent to the unidimensionality assumption is called the local independence assumption. Local independence is that when ability is held constant, an examinee’s responses to any items are statistically independent. Thus, the probability of success on all items for a fixed ability $\theta$ is equal to the product of the separate probabilities of success on each item (Lord, 1980). Let the random variable $U_1, U_2, \ldots, U_j$ take on specific values $u_1, u_2, \ldots, u_J$ ($u_j$ is either 1 or 0), representing the dichotomous responses of examinee to a set of *J* items. Then the local independence assumption is expressed as

$$P\left(U_1 = u_1, U_2 = u_2, \ldots U_n = u_n |\theta\right) = \prod_{j=1}^{J} P_j\left(\theta\right)^{u_j} Q_j\left(\theta\right)^{1-u_j}. \tag{3.2}$$

This is the joint probability of the responses to all *J* items. When unidimensionality is true, local independence can be obtained. In that regard, the two concepts are equivalent (Allen & Yen, 1979). However, local independence can be obtained without satisfying the assumption of unidimensionality. When these assumptions are met, item response theory provides invariant item parameters as well as ability estimates (Hambleton & Swaminathan, 1985).

### Item response function for 2 PL

The two parameter logistic (2PL) model uses two parameters to describe each item, the item discrimination $a_j$ and the item difficulty $b_j$. It allows differently difficult items as well as differently discriminating items. Thus, item characteristic curves vary in slope and location along the ability scale. For a dichotomous item, the item response function is the probability $P_{ij}(\theta)$ of a correct response to the item $j$ for person $i$. It is specified as

$$P_{ij}\left(X_{ij}=1\middle|\theta,a,b\right)=\frac{\exp\left(1.7a_j\left(\theta_i-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i-b_j\right)\right)}. \tag{3.3}$$

The probability 1 – $P_{ij}(\theta)$ or $Q_{ij}(\theta)$ is the chance of person $i$ getting an incorrect response to the item $j$. It is given by

$$1-P_{ij}\left(\theta\right)=Q_{ij}\left(\theta\right)=P\left(X=0\middle|\theta,a,b\right)=\frac{1}{1+\exp\left(1.7a_j\left(\theta_i-b_j\right)\right)}, \tag{3.4}$$

where $\theta_i$ is the ability parameter (*i*= 1, 2, …, *N*), $\theta_i\in\left[-\infty,\infty\right]$, *N* is the number of examinees, and $a_j$ is the item discrimination parameter (*j*= 1, 2, …, *J*) , $a_j\in\left(0,\infty\right]$,which dictates how steeply the ICC rises at its point of maximum discrimination, at the point $\theta_i=b_j$. The usual range for the item discrimination parameter is from 0 to 2. High values of $a_j$ produce the steeper item characteristic functions associated with more discriminating item and low values show a more gradual increase as a function of the ability. *J* is the number of items, $b_j$ is the item difficulty parameter (*j* = 1, 2, …, *J*),

$b_j \in [-\infty, \infty]$, and denotes the point on the ability level at which an examinee has a 0.5 probability of answering item *j* correctly. The value of $b_j$ locates the ICC's inflection point, at which the ICC rises most sharply. Typically, values between -2 to 2 are used for item difficulty parameters. The lower values of $b_j$ are associated with easier items. On the other hand, the higher values of $b_j$ indicate more difficult items. Easy items appear on the left of any item map or ability scale, which means the lower end of the ability scale, and difficult items are the right, or higher end of the ability scale (Hambleton et al., 1991; Parshall et al., 2002).

For item *j*, the conditional distribution for a fixed ability $\theta$ of a single item response is

$$L(u_j|\theta) = \begin{cases} P_j(\theta) & if \quad u_j = 1 \\ Q_j(\theta) & if \quad u_j = 0 \\ 0 & otherwise \end{cases}, \tag{3.5}$$

The local independence assumption is guaranteed by the unidimensionality condition. That is, the probability of a correct response on one item is statistically independent of the probability of correct responses on other items. For a fixed ability level, the joint distribution of all item responses is the product of the distributions of the separate items. Since $\theta_i = (\theta_1, \theta_2, \dots, \theta_N)$, $a_j = (a_1, a_2, \dots, a_J)$, $b_j = (b_1, b_2, \dots, b_J)$, the full likelihood of $U = (u_1, u_2, \dots, u_J)$ for all examinees is

$$L(a,b,\theta|U) \equiv L(a,b,\theta|u_1,u_2,\dots,u_J) = \prod_{i=1}^{N}\prod_{j=1}^{J} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \ . \tag{3.6}$$

The maximum likelihood (ML) estimates are the values of $\theta_i = (\theta_1, \theta_2, \dots, \theta_N)$, $a_j = (a_1, a_2, \dots, a_J)$, $b_j = (b_1, b_2, \dots, b_J)$, that together maximize $L(a,b,\theta|U)$, which are denoted by $\hat{\theta}_i, \hat{a}_j$ and $\hat{b}_j$ respectively. The log likelihood is given by

$$\ln L(a,b,\theta|U) = \sum_{i=1}^{N}\sum_{j=1}^{J}\left[u_{ij}\ln P_{ij} + (1-u_{ij})\ln Q_{ij}\right]. \tag{3.7}$$

The ML estimate of $\left(\theta, a, b\right)$ is denoted by $\left(\hat{\theta}, \hat{a}, \hat{b}\right)$ where $\theta_i = \left(\theta_1, \theta_2, \ldots, \theta_N\right)$, $a_j = \left(a_1, a_2, \ldots, a_J\right)$ and $b_j = \left(b_1, b_2, \ldots, b_J\right)$. Let $l$ define the log-likelihood,

$$
\begin{aligned}
l &= \ln L\left(\theta, a, b \middle| U\right) = \sum_{i=1}^{N}\sum_{j=1}^{J}\left[u_{ij}\ln P_{ij} + \left(1-u_{ij}\right)\ln Q_{ij}\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J}\left[u_{ij}\ln\frac{\exp\left(1.7a_j\left(\theta_i - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - b_j\right)\right)} + \left(1-u_{ij}\right)\ln\frac{1}{1+\exp\left(1.7a_j\left(\theta_i - b_j\right)\right)}\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J}\left[\left(u_{ij}\left(\ln\exp\left(1.7a_j\left(\theta_i - b_j\right)\right) - \ln\left[1+\exp\left(1.7a_j\left(\theta_i - b_j\right)\right)\right]\right)\right)\right. \\
&\quad \left. + \left(1-u_{ij}\right)\left(\ln 1 - \ln\left[1+\exp\left(1.7a_j\left(\theta_i - b_j\right)\right)\right]\right)\right] .
\end{aligned}
\tag{3.8}
$$

Then,

$$
\begin{aligned}
\frac{\partial l}{\partial \theta_k} &= \sum_{j=1}^{J}\left\{\left(u_{kj}\right)\left[\left(1.7a_j\right) - \frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}\right]\right. \\
&\quad \left. + \left(1-u_{kj}\right)\left[-\frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}\right]\right\} \\
&= \sum_{j=1}^{J}\left[\left(1.7a_j\right)u_{kj} - \frac{\left(1.7a_i\right)\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}\right] \\
&= \sum_{j=1}^{J}\left(1.7\right)\left(a_j u_{kj} - a_j\frac{\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}\right) .
\end{aligned}
$$

Let $\frac{dl}{d\theta_k} = 0$. Then the log likelihood is equivalent to

$$
\sum_{j=1}^{J}\left[a_j u_{kj} - a_j\frac{\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}\right] = 0 .
$$

$$\sum_{j=1}^{J} a_j \frac{\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - b_j\right)\right)} = \sum_{j=1}^{J} a_j u_{kj} \text{ , where } k = 1,\ldots,N \text{ .} \tag{3.9}$$

For the 2 PL model, the test score $\sum_{j=1}^{J} a_j u_j$ is a sufficient statistic for estimating examinee ability $\theta_i$ with known item parameters $a_j$. In this case, the test score contains the optimal properties of the MLE. Namely, the estimator of $\theta$ as a function of this also we have sufficient statistic is consistent, efficient and asymptotically normally distributed.

$$\begin{aligned}\frac{\partial l}{\partial b_m} &= \sum_{i=1}^{N}\left\{\left(u_{im}\right)\left[-\left(1.7a_m\right)+\frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}\right]\right.\\ &\left.+\left(1-u_{im}\right)\left[\frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}\right]\right\}\\ &= \sum_{i=1}^{N}\left[-\left(1.7a_m\right)u_{im}+\frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}\right]\\ &= \sum_{i=1}^{N}\left(-1.7\right)\left(a_m u_{im} - a_m\frac{\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}\right).\end{aligned}$$

Let $\frac{dl}{db_m} = 0$. Then the log likelihood estimate obtained from

$$\sum_{i=1}^{N}\left[a_m u_{im} - a_m\frac{\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}\right] = 0 \text{ .}$$

$$\sum_{i=1}^{N} a_m \frac{\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - b_m\right)\right)} = \sum_{i=1}^{N} a_m u_{im} \text{ , where } m = 1,\ldots,J \text{ .} \tag{3.10}$$

$$\frac{\partial l}{\partial a_q} = \sum_{i=1}^{N} \left\{ \left(u_{iq}\right) \left[ \left(1.7\left(\theta_i - b_q\right)\right) - \frac{\left(1.7\left(\theta_i - b_q\right)\right)\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)} \right] \right.$$

$$\left. -\left(1-u_{iq}\right)\left[ \frac{\left(1.7\left(\theta_i - b_q\right)\right)\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - b_q\right)\right)} \right] \right\}$$

$$= \sum_{i=1}^{N} \left[ \left(1.7\left(\theta_i - b_q\right)\right)u_{iq} - \frac{\left(1.7\left(\theta_i - b_q\right)\right)\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)} \right]$$

$$= \sum_{i=1}^{N} (1.7)\left(\theta_i - b_q\right) \left( u_{iq} - \frac{\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)} \right).$$

Let $\frac{dl}{da_q} = 0$. Then the log likelihood estimates of $a$ are obtained from

$$\sum_{i=1}^{N} \left( u_{iq} - \frac{\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i - b_q\right)\right)} \right)\left(\theta_i - b_q\right) = 0, \qquad (3.11)$$

where $q = 1, \ldots, J$. No further simplification is possible.

To estimate the ability parameter and item parameters simultaneously, the joint maximum likelihood (JML) estimation procedure and marginal maximum likelihood (MML) estimation procedure have been used with the 2PL model (Yen & Fitzpatrick, 2006). Although JML estimators have desirable attributes, it has been revealed that such parameter estimates are biased and not consistent. MML is the most commonly used procedure in the field. By avoiding the estimation of the ability parameter, MML procedure improves the estimation of item parameters. Also, an iterative Expectation-Maximization (EM) algorithm is employed. Although the estimators are consistent and asymptotically normal, the statistical properties of the MML estimators have not been convincingly established. Thus, further investigation is recommended (Hambleton & Swaminathan, 1985).

The 2PL IRT model can accommodate a wide variety of real items, which are complex to use and require substantial sample size (Yen & Fitzpatrick, 2006). In particular, the 2PL model can be a valuable model choice for analyzing items with different discrimination as well as different difficulty levels. Thus, it fits to a set of items well and parameters can be estimated accurately (Embretson & Reise, 2000). In addition, it contains a nice statistical property, such as the sufficient statistic, to estimate the ability parameter when item parameters are known. In spite of this, the standard 2 PL IRT model does not account for response time. In fact, when there are time restrictions in the real testing situations, it may not be an appropriate model to use. Therefore, a model is needed that incorporates response time. The 2 PL model has been utilized in the Thissen's (1983) model as a marginal distribution of response accuracy. In the present work, an extended 2PL model is employed as a conditional distribution of response accuracy, which takes response time into consideration.

### An item response model that incorporates response time

The proposed model incorporates response time into a 2 PL model which can be applied to power tests with time limits. By taking response time into account, this model may provide a more realistic description of the item response mechanism than the conventional 2 PL model. The conditional probability of a correct response to item $j$ by examinee $i$ given the response time $t_{ij}$ is

$$f\left(u_{ij}\left|t_{ij},\theta_i,a_j,b_j,r_j,\eta\right.\right)=P_{ij}\left(t_{ij},\theta_i,a_j,b_j,\eta\right)=P_{ij}\left(u_{ij}=1\left|t_{ij},\theta_i,a_j,b_j,\eta\right.\right)=\frac{\exp\left(1.7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)},\tag{3.12}$$

where $\theta_i$ is the ability parameter (*i = 1, 2, …, N*), $a_j$ is the item discrimination parameter (*j = 1, 2, …, J*), $b_j$ is the item difficulty parameter ( *j = 1, 2, …, J*), and $t_{ij}$ is response time by examinee *i* for this particular item *j*. Define $\eta$ as an unknown constant and a regression coefficient for the time variable. If time has no effect on the probability of correct response across items and examinees, then $\eta$ will be zero. On the other hand, if time has an effect on the probability of correct response across items and examinees, then $\eta$ will be non-zero.

## Joint distribution of item response and response time

A joint distribution for item response and response time is proposed in the present work. It is specified as the product of the conditional distribution of response accuracy and the marginal distribution of response time. This is an attempt to overcome the limitation of Thissen's (1983) model where the joint distribution is represented as the product of two marginal distributions, namely response accuracy and response time. The central limitation of his model is due to the inappropriateness of the independence assumption between response accuracy and response time. This makes sense only when response time is not assumed to be independent of person parameters. Therefore, a joint distribution as a product of the conditional distribution of item response and the marginal distribution of response time should provide a more realistic model.

The joint distribution of $u_{ij}$ and $t_{ij}$ is specified as

$$f\left(u_{ij}, t_{ij} \left|\theta_i, s_i, a_j, b_j, r_j\right.\right) = f\left(u_{ij}\left|t_{ij}, \theta_i, a_j, b_j, r_j, \eta\right.\right) f\left(t_{ij}\left|\theta_i, s_i, a_j, b_j, r_j\right.\right). \quad (3.13)$$

The conditional distribution of $u_{ij}$ given $t_{ij}$ is

$$\begin{aligned} f\left(u_{ij}\left|t_{ij}, \theta_i, a_j, b_j, \eta\right.\right) &= P\left(t_{ij}, \theta_i, a_j, b_j, \eta\right)^{u_{ij}} \left[1 - P\left(t_{ij}, \theta_i, a_j, b_j, \eta\right)\right]^{1-u_{ij}} \\ &= \left[\frac{\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}\right]^{u_{ij}} \left[\frac{1}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}\right]^{1-u_{ij}}. \end{aligned} \quad (3.14)$$

As indicated above, if time has no effect on the probability of correct response, then $\eta$ will be zero. Then, the joint distribution model proposed above will reduce to the usual 2PL IRT model.

## Response time model

The distribution of response time is known to be positively skewed and unimodal in general, so that lognormal, Weibull and gamma distributions seem to be the reasonable choices (Schnipke & Scrams, 1999; Verhelst et al., 1997; Wang, 2006). To model the shapes of response time distributions in this study, the lognormal distribution is chosen. This response time distribution captures the idea that the examinees will spend more time on items that correspond to their ability level and spend less time on items either too easy or too difficult. This is consistent with the findings of Wang and Zhang (2006). A log linear model for response times has also been used by Thissen (1983). The lognormal distribution was selected because the parameters are fairly intuitive and thus easier to interpret than parameters of some other distribution (Schnipke & Scrams, 1997). That is to say, the parameters of the lognormal distribution are the mean and standard deviation of the natural logarithm of the original values in reference to a normal distribution (Casella & Berger, 2002). Thus, parameters can be easily estimated from sample statistics (Schnipke & Scrams, 1999). In addition, the lognormal model has been empirically tested against models based on the normal, gamma and Weibull distributions. It has been shown an excellent fit to the response times and outperformed other distributions that have been studied (Schnipke & Scrams, 1999; Storms & Delbeke, 1992; van der Linden et al., 1999; van der Linden et al., 2003; van der Linden, 2006). In fact, Schnipke and Scrams (1999) report that the lognormal distribution not only fits the best of all but also provides a very good fit for most items.

In the lognormal distribution, the scale parameter $\mu$ and the shape parameter $\sigma$ determine the skewness of the distribution. If the response time density, $f\left(t\right)$, has a lognormal distribution, then the natural logarithm of time $t$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$.

In Figure 3.1, the probability density functions all start at zero, increases to their mode, and decrease thereafter. For a fixed $\mu$, the degree of skewness increases as $\sigma$ increases. For values greater than 1, e.g., $\sigma$=10, the *pdf* rises very sharply in the beginning and essentially follows the ordinate axis, peaks out, and then decreases sharply. In Figure 3.2, for a fixed $\sigma$, as $\mu$ increases the skewness of the probability density function decreases.

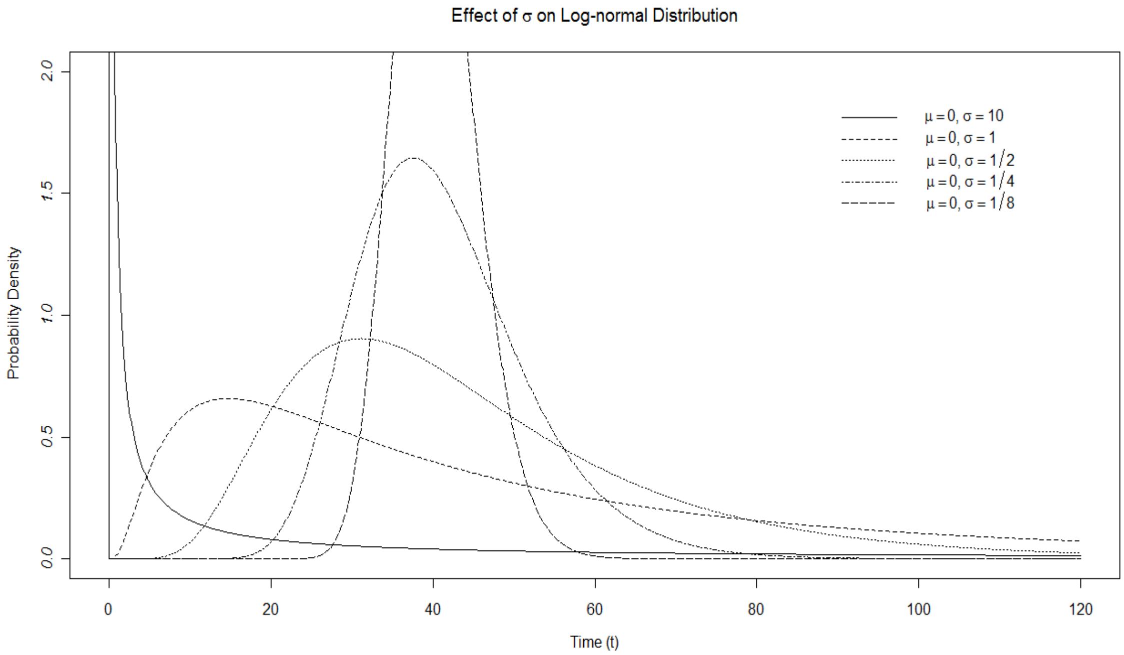


Figure 3.1 The plot of the lognormal probability density function for five values of $\sigma$

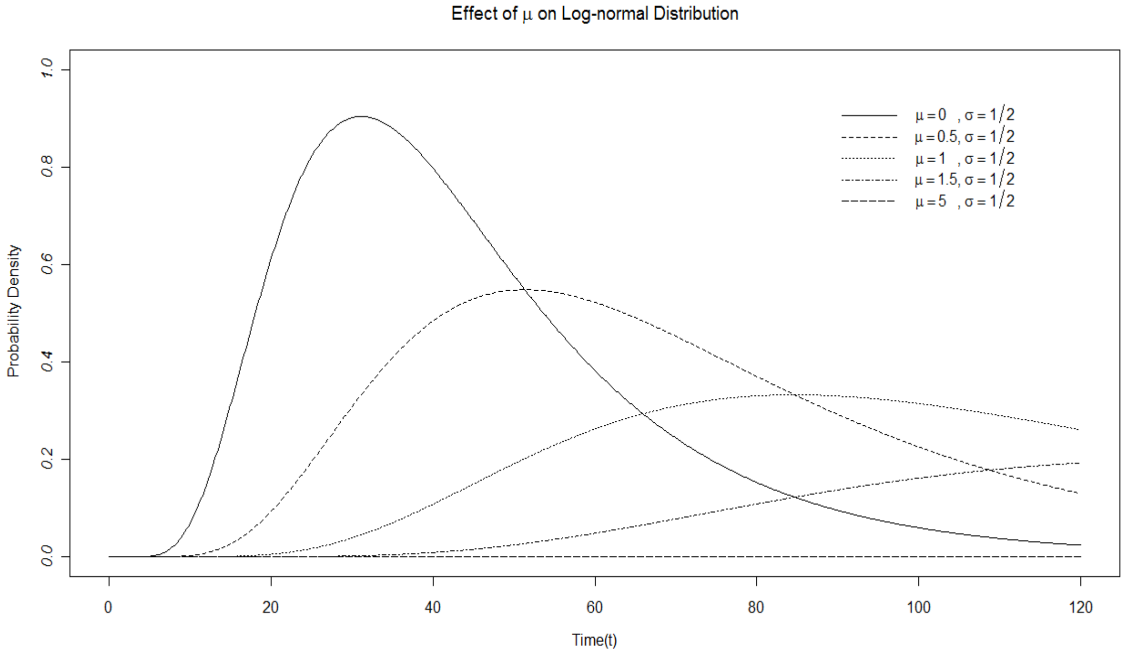


Figure 3.2 The plot of the lognormal probability density function for five values of $\mu$

In the present work, a modified version of Thissen's (1983) log linear model will be employed to fit the response time. It is expressed as

$$\ln t_{ij} = \upsilon + s_i + r_j + gz_{ij} + \varepsilon_{ij} \,, \tag{3.15}$$

where $z_{ij} = 1.7a_j\left(\theta_i - b_j\right)$, $\upsilon$ is overall mean indicating the general response time required by item $j$, $\upsilon$ is assumed to be zero. $s_i$ is person slowness parameter which exhibits the slowness of examinee $i$. Next, $r_j$ is an item slowness parameter. It is a parameter for the response time required by item $j$. Finally, $g$ is an unknown constant and describes the log-linear relationship between response time and examinee ability. Originally, the log-linear relationship between response time and examinee ability defined as a negative relationship in Thissen (1983). He assumed a monotonically decreasing relation between slowness and ability by stating that "more able students work faster" (Thissen, 1983, p. 202). However, van der Linden and van Krimpen-Stoop (2003) found rather a monotonically increasing relation between slowness and ability. If an examinee selects accuracy rather than speed, both the value of ability and slowness parameter seem to increase. Also, empirical evidence proved that the relation between ability and slowness depended on the level of speededness of the test (van der Linden & van Krimpen-Stoop, 2003). Therefore, unlike Thissen's (1983) original version, a positive sign is used to describe a more general relationship between response time and examinee ability in this model. $\varepsilon_{ij}$ is a normally distributed residual with mean 0 and variance $\sigma^2$, that is, $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$.

Lognormal marginal distribution for response time is expressed as

$$\ln t_{ij} \sim N\left(\upsilon + s_i + r_j + gz_{ij}, \sigma^2\right)$$

$$f\left(t_{ij} \middle| \theta_i, a_j, b_j, s_i, r_j\right) = \frac{1}{\sqrt{2\pi}\sigma t_{ij}} \exp\left(\frac{-\left[\ln t_{ij} - \left(\upsilon + s_i + r_j + gz_{ij}\right)\right]^2}{2\sigma^2}\right), \tag{3.16}$$

where $\upsilon + s_i + r_j + gz_{ij}$ is the scale parameter, namely, the mean of the log response time expressed in log seconds, and $\sigma$ is a shape parameter, that is, the standard deviation of the log response time expressed in log seconds.

**Joint likelihood of the item responses and the response times**

The joint likelihood of the item responses and response times is given by

$$\begin{aligned}
&L\left(\theta, a, b, \eta \middle| U, T\right) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} f\left(u_{ij}, t_{ij} \middle| \theta_i, s_i, a_j, b_j, r_j\right) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} f\left(u_{ij} \middle| t_{ij}, \theta_i, a_j, b_j, r_j, \eta\right) f\left(t_{ij} \middle| \theta_i, s_i, a_j, b_j, r_j\right) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} \left[ \frac{\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right]^{u_{ij}} \left[ \frac{1}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right]^{1-u_{ij}} \\
&\left[ \frac{1}{\sqrt{2\pi}\sigma t_{ij}} \exp\left( \frac{-\left[\ln t_{ij} - \left(\upsilon + s_i + r_j + g z_{ij}\right)\right]^2}{2\sigma^2} \right) \right] .
\end{aligned} \tag{3.17}$$

The log-likelihood equation is

$$l\left(\theta, a, b, \eta \middle| U, T\right) = \sum_{i=1}^{N} \sum_{j=1}^{J} \ln\left[ f\left(u_{ij}, t_{ij} \middle| \theta_i, s_i, a_j, b_j, r_j\right) \right]$$

$$\begin{aligned}
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg[ u_{ij} \ln \frac{\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} + \left(1-u_{ij}\right) \ln \frac{1}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \\
&+ \ln\left(2\pi\right)^{-\frac{1}{2}} + \ln\left(\sigma\right)^{-1} + \ln\left(t_{ij}\right)^{-1} + \ln \exp\left( -\frac{\left[\ln t_{ij} - \left(\upsilon + s_i + r_j + g\left(1.7a_j\left(\theta_i - b_j\right)\right)\right)\right]^2}{2\sigma^2} \right) \Bigg] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg[ \left( u_{ij} \left( \ln \exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right) - \ln\left[1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)\right] \right) \right) \\
&+ \left(1-u_{ij}\right)\left( \ln 1 - \ln\left[1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)\right] \right) \\
&- \frac{1}{2} \ln 2\pi - \ln \sigma - \ln t_{ij} + \ln \exp\left( -\frac{\left[\ln t_{ij} - \left(\upsilon + s_i + r_j + g\left(1.7a_j\left(\theta_i - b_j\right)\right)\right)\right]^2}{2\sigma^2} \right) \Bigg] .
\end{aligned} \tag{3.18}$$

Taking the derivative with respect to $\theta_k$, it follows that

$$\frac{\partial l}{\partial \theta_k} = \sum_{j=1}^{J} \left\{ \left(u_{kj}\right) \left[ \left(1.7a_j\right) - \frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} \right] \right.$$

$$\left. + \left(1-u_{kj}\right) \left[ -\frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} \right] \left( \frac{2\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]\left(g1.7a_j\right)}{2\sigma^2} \right) \right\}$$

$$= \sum_{j=1}^{J} \left[ \left(1.7a_j\right)u_{kj} - \frac{\left(1.7a_i\right)\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} + \frac{\left(g1.7a_j\right)\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]}{\sigma^2} \right]$$

$$= \sum_{j=1}^{J} \left(1.7a_j\right) \left( u_{kj} - \frac{\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} + \frac{g\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]}{\sigma^2} \right).$$

Let $\dfrac{dl}{d\theta_k} = 0$. Then the likelihood equations are

$$\sum_{j=1}^{J} \left(a_j\right) \left( u_{kj} - \frac{\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} + \frac{g\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]}{\sigma^2} \right) = 0,$$

(3.19)

where $k = 1, \ldots, N$. No further simplification is possible. Next, we consider the estimator of $a_j$. The partial derivative is

$$
\begin{aligned}
\frac{\partial l}{\partial a_q} &= \sum_{i=1}^{N}\left\{\left(u_{iq}\right)\left[\left(1.7\left(\theta_i-\eta t_{iq}-b_q\right)\right)-\frac{\left(1.7\left(\theta_i-\eta t_{iq}-b_q\right)\right)\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}\right]\right.\\
&-\left(1-u_{iq}\right)\left[\frac{\left(1.7\left(\theta_i-\eta t_{iq}-b_q\right)\right)\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i-\eta t_{iq}-b_q\right)\right)}\right]\\
&\left.+\frac{2\left[\ln t_{iq}-\left(\upsilon+s_i+r_q+g\left(1.7a_q\left(\theta_i-b_q\right)\right)\right)\right]\left(g1.7\left(\theta_i-b_q\right)\right)}{2\sigma^2}\right\}\\
&=\sum_{i=1}^{N}\left\{\left(1.7\left(\theta_i-\eta t_{iq}-b_q\right)\right)u_{iq}-\frac{\left(1.7\left(\theta_i-\eta t_{iq}-b_q\right)\right)\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}\right.\\
&\left.+\frac{\left[\ln t_{iq}-\left(\upsilon+s_i+r_q+g\left(1.7a_q\left(\theta_i-b_q\right)\right)\right)\right]\left(g1.7\left(\theta_i-b_q\right)\right)}{\sigma^2}\right\}\\
&=\sum_{i=1}^{N}(1.7)\left\{\left(\theta_i-\eta t_{iq}-b_q\right)u_{iq}-\frac{\left(\theta_i-\eta t_{iq}-b_q\right)\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}\right.\\
&\left.+\frac{\left(g\left(\theta_i-b_q\right)\right)\left[\ln t_{iq}-\left(\upsilon+s_i+r_q+g\left(1.7a_q\left(\theta_i-b_q\right)\right)\right)\right]}{\sigma^2}\right\}.
\end{aligned}
$$

Let $\dfrac{dl}{da_q}=0$. Then the likelihood equations are

$$
\sum_{i=1}^{N}\left(\left(\theta_i-\eta t_{iq}-b_q\right)u_{iq}-\frac{\left(\theta_i-\eta t_{iq}-b_q\right)\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}{1+\exp\left(1.7a_q\left(\theta_i-\eta t_{iq}-b_q\right)\right)}+\frac{\left(g\left(\theta_i-b_q\right)\right)\left[\ln t_{iq}-\left(\upsilon+s_i+r_q+g\left(1.7a_q\left(\theta_i-b_q\right)\right)\right)\right]}{\sigma^2}\right)=0. \tag{3.20}
$$

where $q=1,\ldots,J$. No further simplification is possible. Next, for the difficulty $b_j$, we have

$$
\begin{aligned}
\frac{\partial l}{\partial b_m} &= \sum_{i=1}^{N} \left\{ \left(u_{im}\right) \left[ -\left(1.7a_m\right) + \frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)} \right] \right. \\
&\left. + \left(1-u_{im}\right) \left[ \frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)} \right] - \frac{2\left[\ln t_{im} - \left(\upsilon + s_i + r_m + g\left(1.7a_m\left(\theta_i - b_m\right)\right)\right)\right]\left(g1.7a_m\right)}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{N} \left\{ -\left(1.7a_m\right)u_{im} + \frac{\left(1.7a_m\right)\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)} \right. \\
&\left. - \frac{\left(g1.7a_m\right)\left[\ln t_{im} - \left(\upsilon + s_i + r_m + g\left(1.7a_m\left(\theta_i - b_m\right)\right)\right)\right]}{\sigma^2} \right\} \\
&= \sum_{i=1}^{N} \left(-1.7a_m\right) \left( u_{im} - \frac{\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)} + \frac{g\left[\ln t_{im} - \left(\upsilon + s_i + r_m + g\left(1.7a_m\left(\theta_i - b_m\right)\right)\right)\right]}{\sigma^2} \right).
\end{aligned}
$$

Let $\frac{dl}{db_m} = 0$. Then the log likelihood is equivalent to

$$
\sum_{i=1}^{N} \left(a_m\right) \left( u_{im} - \frac{\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)}{1+\exp\left(1.7a_m\left(\theta_i - \eta t_{im} - b_m\right)\right)} + \frac{g\left[\ln t_{im} - \left(\upsilon + s_i + r_m + g\left(1.7a_m\left(\theta_i - b_m\right)\right)\right)\right]}{\sigma^2} \right) = 0, \tag{3.21}
$$

where $m=1,\ldots,J$. No further simplification is possible.

$$
\begin{aligned}
\frac{\partial l}{\partial \eta} &= \sum_{i=1}^{N} \sum_{j=1}^{J} \left\{ \left(u_{ij}\right) \left[ -\left(1.7a_j t_{ij}\right) + \frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right] \right. \\
&\left. + \left(1-u_{ij}\right) \left[ \frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right] \right\} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \left[ -\left(1.7a_j t_{ij}\right)u_{ij} + \frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \left(-1.7a_j t_{ij}\right) \left( u_{ij} - \frac{\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i - \eta t_{ij} - b_j\right)\right)} \right).
\end{aligned}
$$

Let $\frac{dl}{d\eta}=0$. Then the log likelihood is equivalent to

$$\sum_{i=1}^{N}\sum_{j=1}^{J}\left(a_j t_{ij}\right)\left(u_{ij}-\frac{\exp\left(1.7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_i-\eta t_{ij}-b_j\right)\right)}\right)=0\,. \tag{3.22}$$

No further simplification is possible.

$$\begin{aligned}\frac{\partial l}{\partial s_x}&=\sum_{j=1}^{J}\left[\frac{2\left[\ln t_{xj}-\left(\upsilon+s_x+r_j+g\left(1.7a_j\left(\theta_x-b_j\right)\right)\right)\right]}{2\sigma^2}\right]\\&=\sum_{j=1}^{J}\left[\frac{\left[\ln t_{xj}-\left(\upsilon+s_x+r_j+g\left(1.7a_j\left(\theta_x-b_j\right)\right)\right)\right]}{\sigma^2}\right].\end{aligned}$$

Let $\frac{dl}{ds_x}=0$. Then the log likelihood is equivalent to

$$\sum_{j=1}^{J}\frac{\ln t_{xj}-\left(\upsilon+s_x+r_j+g\left(1.7a_j\left(\theta_x-b_j\right)\right)\right)}{\sigma^2}=0\,, \tag{3.23}$$

where $x=1,\ldots,N$ . No further simplification is possible.

$$\begin{aligned}\frac{\partial l}{\partial r_y}&=\sum_{i=1}^{N}\frac{2\left[\ln t_{iy}-\left(\upsilon+s_i+r_y+g\left(1.7a_y\left(\theta_i-b_y\right)\right)\right)\right]}{2\sigma^2}\\&=\sum_{i=1}^{N}\frac{\left[\ln t_{iy}-\left(\upsilon+s_i+r_y+g\left(1.7a_y\left(\theta_i-b_y\right)\right)\right)\right]}{\sigma^2}\,.\end{aligned}$$

Let $\frac{dl}{dr_y}=0$. Then the log likelihood is equivalent to

$$\sum_{i=1}^{N}\frac{\ln t_{iy}-\left(\upsilon+s_i+r_y+g\left(1.7a_y\left(\theta_i-b_y\right)\right)\right)}{\sigma^2}=0\,, \tag{3.24}$$

where $y=1,\ldots,J$ . No further simplification is possible.

$$\frac{\partial l}{\partial g}=\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{2\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]\left(1.7a_j\left(\theta_i-b_j\right)\right)}{2\sigma^2}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{\left(1.7a_j\left(\theta_i-b_j\right)\right)\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]}{\sigma^2}.$$

Let $\frac{dl}{dg}=0$. Then the log likelihood is equivalent to

$$\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{\left(1.7a_j\left(\theta_i-b_j\right)\right)\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]}{\sigma^2}=0. \qquad (3.25)$$

No further simplification is possible.

$$\frac{\partial l}{\partial \sigma}=\sum_{i=1}^{N}\sum_{j=1}^{J}\left[-\frac{1}{\sigma}+\frac{2\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]^2}{2\sigma^3}\right]$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{J}\left[-\frac{1}{\sigma}+\frac{\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]^2}{\sigma^3}\right]$$

$$=\left(-\frac{1}{\sigma}\right)\left[NJ-\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]^2}{\sigma^2}\right].$$

Let $\frac{dl}{d\sigma}=0$. Then the log likelihood is equivalent to

$$\sigma^2=\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left[\ln t_{ij}-\left(\upsilon+s_i+r_j+g\left(1.7a_j\left(\theta_i-b_j\right)\right)\right)\right]^2}{NJ}. \qquad (3.26)$$

This is a closed form solution.

### Marginal maximum likelihood estimation

The simultaneous parameter estimation in the typical 2PL item response theory (IRT) model using joint maximum likelihood (JML) has been shown to be biased and not consistent (Hambleton & Swaminathan, 1985; Yen & Fitzpatrick, 2006). Thus, the original motivating model, joint maximum likelihood (JML) (See equation 3.17), as an extension of the item response model, leads to estimates with inconsistent and improper statistical properties. However, marginal maximum likelihood (MML) estimation using an iterative Expectation-Maximization (EM) algorithm has been proven to provide estimates that are consistent and asymptotically normal (Bock & Aitkin, 1981). Consequently, it has been the most commonly used procedure in the field. Therefore, MML estimation is derived in this work.

Among MML estimates of item parameters for various IRT models based on EM algorithm, those of Bock and Aitkin (1981) and Woodruff and Hanson (1997) will be compared with the MML estimation given in this work. That is to say, Bock and Aitkin's (1981) algorithm is most commonly utilized in IRT models (van der Linden & Hambleton, 1997). Also, Woodruff and Hanson's (1997) algorithm has been utilized in the extended IRT models, such as the models of Wang and Hanson (2005) and Wang (2006), which include two person parameters and response time.

The MML estimates of item parameters using EM algorithm have shown to utilize different approximation approaches. Specifically, Bock and Aitkin (1981) derived estimation procedures based on a continuous latent variable. Then, for computational purposes, EM algorithm was used to implement approximations of those procedures with a discrete version of the continuous latent variable. To be precise, instead of computing the likelihood, the likelihood was maximized individually by a simpler likelihood. In order to reduce the computational demands, the approximation was done by replacing the latent variable by its conditional expectation, namely, by sorting response vectors into item score patterns within the same score groups (Bock & Aitkin, 1981). Woodruff and Hanson (1997) specified their model by approximating a continuous latent variable with a discrete latent variable. The EM algorithm for finding maximum likelihood estimates of finite mixtures was applied to this discrete latent variable. As in a commonly applied EM procedure, values of the latent variables were fixed beforehand and the probabilities of those values were allowed to vary. (Those probabilities are the parameters.) In the E-step, the expectation of the complete data likelihood was computed. Then, in the M-step, the expectation of the complete data likelihood was maximized over the set of item parameters and the probabilities. That is to say, instead of calculating observed data likelihood, the expected completed data was used to maximize the complete likelihood (Woodruff & Hanson, 1997).

The approximation approach by discretizing the distributions for latent variables presented in this work is considerably different than the likelihood approximation obtained by Bock and Aitkin (1981) and Woodruff and Hanson (1997). When Woodruff and Hanson (1997) discretized the latent variable, no attempt was made for discretized latent variable to look like the normal distribution. In particular, if we require the latent variable to take ten arbitrary values with different probabilities, there is no way to make this look like a normal distribution. If you assume ability to have a normal distribution, then the Woodruff and Hanson (1997) algorithm does not guarantee this, because their method does not force the distribution. However, if the latent variable is allowed to be approximately normal, then the method presented here achieves this by using a discrete distribution whose values are spaced to resemble a normal distribution. In other words, the model approximated the normal distribution for ability parameter $\theta$ by having $K$ values with equal probability of $1/K$, with spacing chosen to resemble the normal distribution. The same is done for the slowness parameter $s$ using $L$ values with equal probability of $1/L$. This discrete distribution may be regarded as an idealized sample from the normal distribution.

In addition, it is noticed that the EM way of computing is very slow. The marginal maximum likelihood approach presented here converges faster than EM algorithm because the direct likelihood is easy to compute due to the discretizing of the latent variables. That is, this approach uses a simple sum as suppose to integral because we have discretized the distribution. By using the discrete normal distributions for the latent variables, a computable genuine marginal maximum likelihood (Equation 3.27) is used, instead of the EM algorithm. Hence, the likelihood function is actually computed and maximized directly.

In summary, the approach presented in this work is unique. Instead of approximating integrals as in the EM algorithm, the distributions of the latent variables are approximated by using idealized spacing to resemble the normal distribution. Then, a proceeding procedure using marginal maximum likelihood incorporates this discrete distribution. As a result, a genuine marginal maximum likelihood is computed and maximized directly which leads to faster convergence than EM procedure.

In the model presented (Equation 3.27), the ability parameter $\theta$ and the person slowness parameter $s$ are two latent variables which are taken to be discrete. In addition, the MML estimation is derived based on the discrete latent variables. The model is specified as follows: The ability parameter $\theta \in \left\{\sigma_q q_1, \sigma_q q_2, \ldots, \sigma_q q_K\right\}$ is assumed to be from discrete $N\left(0, \sigma_q\right)$ where $\sigma_q$ is an unknown scale factor. The values $q_1, \ldots, q_K$ are chosen to resemble a standard normal distribution ($N\left(0,1\right)$). The person slowness parameter $s \in \left\{\sigma_s s_1, \sigma_s s_2, \ldots, \sigma_s s_L\right\}$ is assumed to be from discrete $N\left(0, \sigma_s\right)$ where $\sigma_s$ is unknown. The values $s_1, \ldots, s_L$ are known and chosen to resemble a $N\left(0,1\right)$ distribution. It is assumed that $\theta$ and $s$ are independently distributed. The two latent variables, $\theta$ and $s$, are integrated out by summing over the discrete distribution of *K* and *L* spaced values. It leads the marginal maximum likelihood to only involve with observed data $u_{ij}$ and $t_{ij}$, where $u_{ij}$ is expressed as the item response accuracy by examinee *i* for the item *j* and $t_{ij}$ is response time by examinee *i* for item *j*. Marginal maximum likelihood of the item responses and the response times is given by

$$
\begin{aligned}
&L\left(\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta \middle| U,T\right)\\
&=\prod_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J} f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)\\
&=\prod_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J} f\left(u_{ij}\middle|t_{ij},\sigma_q q_k,a_j,b_j,r_j,\eta\right) f\left(t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right)\\
&=\prod_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}\left[\frac{\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}\right]^{u_{ij}}\left[\frac{1}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}\right]^{1-u_{ij}}\\
&\left[\frac{1}{\sqrt{2\pi}\sigma t_{ij}}\exp\left(\frac{-\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+gz_{ij}\right)\right]^2}{2\sigma^2}\right)\right].
\end{aligned}
\tag{3.27}
$$

where $a_j$ is the item discrimination parameter ($j$ = 1, 2, …, $J$), $b_j$ is the item difficulty parameter ( $j$ = 1, 2, …, $J$). Define $\eta$ as an unknown constant and a regression coefficient for the time variable. Define $z_{ij}=1.7a_j\left(\theta_i-b_j\right)$, $\upsilon$ is overall mean indicating the general response time required by item $j$, $s_i$ is the person slowness parameter which exhibits the slowness of examinee $i$. Next, $r_j$ is an item slowness parameter. It is a parameter that represents the time needed to respond to item $j$, and $g$ is an unknown constant that describes the log-linear relationship between response time and examinee ability. Finally, $\varepsilon_{ij}$ is a normally distributed residual with mean 0 and variance $\sigma^2$, that is, $\varepsilon_{ij}\sim N\left(0,\sigma^2\right)$. Here, $\upsilon+\sigma_s s_l+r_j+gz_{ij}$ is the scale parameter, namely, the mean of the log response time expressed in log seconds, and $\sigma$ is a shape parameter, that is, the standard deviation of the log response time expressed in log seconds.

The log-likelihood equation is

$$
l\left(\theta,a,b,\eta\middle|U,T\right)=\sum_{i=1}^{N}\ln\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}\left[f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)\right].
\tag{3.28}
$$

The partial derivative with respect to $a_{j'}$ is

$$\frac{\partial l}{\partial a_{j'}}$$

$$= \sum_{i=1}^{N} \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\frac{\partial}{\partial a_{j'}}\ln f\left(u_{ij'}, t_{ij'}\left|\sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right.\right)\prod_{j=1}^{J} f\left(u_{ij}, t_{ij}\left|\sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J} f\left(u_{ij}, t_{ij}\left|\sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right.\right)}$$

$$= 0, \tag{3.29}$$

where

$$\ln f\left(u_{ij}, t_{ij}\left|\sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right.\right)$$

$$= u_{ij}\ln\frac{\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right)} + \left(1-u_{ij}\right)\ln\frac{1}{1+\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right)}$$

$$+\ln\left(2\pi\right)^{-1/2} + \ln\left(\sigma\right)^{-1} + \ln\left(t_{ij}\right)^{-1} + \ln\exp\left(-\frac{\left[\ln t_{ij} - \left(\upsilon + \sigma_s s_l + r_j + g\left(1.7a_j\left(\sigma_q q_k - b_j\right)\right)\right)\right]^2}{2\sigma^2}\right)$$

$$= \left(u_{ij}\left(\ln\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right) - \ln\left[1+\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right)\right]\right)\right)$$

$$+\left(1-u_{ij}\right)\left(\ln 1 - \ln\left[1+\exp\left(1.7a_j\left(\sigma_q q_k - \eta t_{ij} - b_j\right)\right)\right]\right)$$

$$-\frac{1}{2}\ln 2\pi - \ln\sigma - \ln t_{ij} + \left(-\frac{\left[\ln t_{ij} - \left(\upsilon + \sigma_s s_l + r_j + g\left(1.7a_j\left(\sigma_q q_k - b_j\right)\right)\right)\right]^2}{2\sigma^2}\right).$$

Then

$$\begin{aligned}
&\frac{\partial}{\partial a_{j'}} \ln f\left(u_{ij'}, t_{ij'} \middle| \sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right) \\
&= \left(u_{ij'}\right)\left[\left(1.7\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right) - \frac{\left(1.7\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)\exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1 + \exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}\right] \\
&- \left(1 - u_{ij'}\right)\left[\frac{\left(1.7\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)\exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1 + \exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}\right] \\
&+ \frac{2\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7 a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]\left(g 1.7\left(\sigma_q q_k - b_{j'}\right)\right)}{2\sigma^2} \\
&= \left(1.7\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right) u_{ij'} - \frac{\left(1.7\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)\exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1 + \exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)} \\
&+ \frac{\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7 a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]\left(g 1.7\left(\sigma_q q_k - b_{j'}\right)\right)}{\sigma^2} \\
&= (1.7)\left\{\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right) u_{i_{j'}} - \frac{\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1 + \exp\left(1.7 a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}\right. \\
&+ \frac{\left(g\left(\sigma_q q_k - b_{j'}\right)\right)\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7 a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]}{\sigma^2} \; .
\end{aligned}$$

Also,

$$\begin{aligned}
&\frac{\partial l}{\partial b_{j'}} \\
&= \sum_{i=1}^{N} \frac{\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{1}{K}\frac{1}{L}\frac{\partial}{\partial b_{j'}} \ln f\left(u_{ij'}, t_{ij'} \middle| \sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right) \prod_{j=1}^{J} f\left(u_{ij}, t_{ij} \middle| \sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{1}{K}\frac{1}{L} \prod_{j=1}^{J} f\left(u_{ij}, t_{ij} \middle| \sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right)} \\
&= 0,
\end{aligned} \tag{3.30}$$

where

$$
\begin{aligned}
&\frac{\partial}{\partial b_{j'}} \ln f\left(u_{ij'}, t_{ij'} \middle| \sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right) \\
&= \left(u_{ij'}\right)\left[-\left(1.7a_{j'}\right) + \frac{\left(1.7a_{j'}\right)\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1+\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}\right] \\
&+\left(1-u_{ij'}\right)\left[\frac{\left(1.7a_{j'}\right)\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1+\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}\right] \\
&-\frac{2\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]\left(g1.7a_{j'}\right)}{2\sigma^2} \\
&= -\left(1.7a_{j'}\right)u_{ij'} + \frac{\left(1.7a_{j'}\right)\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1+\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)} \\
&-\frac{\left(g1.7a_{j'}\right)\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]}{\sigma^2} \\
&= \left(-1.7a_{j'}\right)\left(u_{ij'} - \frac{\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)}{1+\exp\left(1.7a_{j'}\left(\sigma_q q_k - \eta t_{ij'} - b_{j'}\right)\right)} + \frac{g\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]}{\sigma^2}\right).
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial l}{\partial r_{j'}} \\
&= \sum_{i=1}^{N} \frac{\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{1}{K}\frac{1}{L}\frac{\partial}{\partial r_{j'}} \ln f\left(u_{ij'}, t_{ij'} \middle| \sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right) \prod_{j=1}^{J} f\left(u_{ij}, t_{ij} \middle| \sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{1}{K}\frac{1}{L} \prod_{j=1}^{J} f\left(u_{ij}, t_{ij} \middle| \sigma_q q_k, \sigma_s s_l, a_j, b_j, r_j, \eta\right)} \\
&= 0,
\end{aligned}
\tag{3.31}
$$

where

$$
\begin{aligned}
&\frac{\partial}{\partial r_{j'}} \ln f\left(u_{ij'}, t_{ij'} \middle| \sigma_q q_k, \sigma_s s_l, a_{j'}, b_{j'}, r_{j'}, \eta\right) \\
&= \frac{2\left[\ln t_{i_{j'}} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]}{2\sigma^2} \\
&= \frac{\left[\ln t_{ij'} - \left(\upsilon + \sigma_s s_l + r_{j'} + g\left(1.7a_{j'}\left(\sigma_q q_k - b_{j'}\right)\right)\right)\right]}{\sigma^2}.
\end{aligned}
$$

$$\begin{aligned}
&\frac{\partial l}{\partial \eta} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\left\{\sum_{j=1}^{J}\frac{\partial}{\partial \eta}\ln f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)\right\}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\sum_{j=1}^{J}\frac{\partial}{\partial \eta}\ln f\left(u_{ij}\middle|t_{ij},\sigma_q q_k,a_j,b_j,r_j,\eta\right)\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\middle|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right)} \\
&=0,
\end{aligned} \tag{3.32}$$

where

$$\begin{aligned}
&\frac{\partial}{\partial \eta}\ln f\left(u_{ij}\middle|t_{ij},\sigma_q q_k,a_j,b_j,r_j,\eta\right) \\
&=\left(u_{ij}\right)\left[-\left(1.7a_j t_{ij}\right)+\frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}\right] \\
&+\left(1-u_{ij}\right)\left[\frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}\right] \\
&=-\left(1.7a_j t_{ij}\right)u_{ij}+\frac{\left(1.7a_j t_{ij}\right)\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)} \\
&=\left(-1.7a_j t_{ij}\right)\left(u_{ij}-\frac{\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\sigma_q q_k-\eta t_{ij}-b_j\right)\right)}\right)\ .
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial l}{\partial g} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\left\{\sum_{j=1}^{J}\frac{\partial}{\partial g}\ln f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)\right\}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\sum_{j=1}^{J}\frac{\partial}{\partial g}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right)\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)} \\
&=0,
\end{aligned}
\tag{3.33}
$$

where

$$
\begin{aligned}
&\frac{\partial}{\partial g}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right) \\
&=\frac{2\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)}{2\sigma^2} \\
&=\frac{\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]}{\sigma^2}.
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial l}{\partial \sigma} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\left\{\sum_{j=1}^{J}\frac{\partial}{\partial \sigma}\ln f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)\right\}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)} \\
&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\sum_{j=1}^{J}\frac{\partial}{\partial \sigma}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right)\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)} \\
&=0,
\end{aligned}
\tag{3.34}
$$

where

$$\begin{aligned}&\frac{\partial}{\partial\sigma}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right)\\&=-\frac{1}{\sigma}+\frac{2\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]^2}{2\sigma^3}\\&=-\frac{1}{\sigma}+\frac{\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]^2}{\sigma^3}\ .\end{aligned}$$

$$\begin{aligned}&\frac{\partial l}{\partial\sigma_s}\\&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\left\{\sum_{j=1}^{J}\frac{\partial}{\partial\sigma_s}\ln f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)\right\}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}\\&=\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\sum_{j=1}^{J}\frac{\partial}{\partial\sigma_s}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right)\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}{\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{K}\frac{1}{L}\prod_{j=1}^{J}f\left(u_{ij},t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j,\eta\right.\right)}\\&=0,\end{aligned}\tag{3.35}$$

where

$$\begin{aligned}&\frac{\partial}{\partial\sigma_s}\ln f\left(t_{ij}\left|\sigma_q q_k,\sigma_s s_l,a_j,b_j,r_j\right.\right)\\&=\frac{2s_l\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]}{2\sigma^2}\\&=\frac{s_l\left[\ln t_{ij}-\left(\upsilon+\sigma_s s_l+r_j+g\left(1.7a_j\left(\sigma_q q_k-b_j\right)\right)\right)\right]}{\sigma^2}\ .\end{aligned}$$

For the estimation procedure, the *BFGS* (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) method is used. It is a numerical quasi-Newton algorithm. The likelihood is optimized by a general optimization procedure. Namely, the *BFGS* method uses repeated evaluation of function and gradient at various points, then finds the maximum of the likelihood function in an efficient way. In R program, *BFGS* method has been implemented as one of the input options to the *optim* function (R Development Core Team, 2007).

### Maximum a posteriori estimation

Maximum a posteriori (MAP) estimation is employed to estimate person parameters. This Baysian estimation approach is an attempt to overcome some of the limitations associated with the joint maximum likelihood (JML) estimation, specifically, the failure to estimate ability levels of examinees with all correct or all incorrect responses because the likelihood goes to infinity.

The original motivating model, joint maximum likelihood (JML) (Equation 3.17), as an extension of the item response model, produces inconsistent and improper statistical properties. Many researchers in the measurement area have indicated that this limitation of JML estimation can be handled by incorporating prior information, with methods such as MAP estimation. Namely, the prior distribution of ability parameter values is used in conjunction with the log-likelihood function to derive an ability level estimate, based on maximizing the posterior distribution. It has been shown that for test lengths longer than 20 items, the prior distribution is swamped by the likelihood function and has no effect on ability parameter estimation (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). As a result, the MAP estimation of the ability parameters $(\theta)$ and person slowness parameters $(s)$ is applied in this work.

The objective of MAP estimation is to find the values of ability parameters $(\theta)$ and person slowness parameters $(s)$ that maximize the posterior distribution. The prior distribution assumes that $\theta$ and $s$ are independent with $\theta \sim N(0,1)$ and $s \sim N(0,\sigma_s^2)$ respectively. The posterior distribution is proportional to the product of prior and likelihood over one examinee:

$$\prod_{j=1}^{J} f\left(u_j,t_j\middle|\theta,s,a_j,b_j,r_j,\eta\right)g\left(\theta\right)h\left(s\right)$$

$$=\prod_{j=1}^{J} f\left(u_j\middle|t_j,\theta,a_j,b_j,r_j,\eta\right)f\left(t_j\middle|\theta,s,a_j,b_j,r_j\right)g\left(\theta\right)h\left(s\right)$$

$$=\prod_{j=1}^{J}\left[\frac{\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}\right]^{u_j}\left[\frac{1}{1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}\right]^{1-u_j}$$

$$\left[\frac{1}{\sqrt{2\pi}\sigma t_j}\exp\left(\frac{-\left[\ln t_j-\left(\upsilon+s+r_j+gz_j\right)\right]^2}{2\sigma^2}\right)\right]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\theta^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}\sigma_s}\exp\left(-\frac{s^2}{2\sigma_s^2}\right)\right].$$

(3.36)

Or equivalently maximizing the logarithm here, as

$$\sum_{j=1}^{J}\ln f\left(u_j,t_j\middle|\theta,s,a_j,b_j,r_j,\eta\right)+\ln g\left(\theta\right)+\ln h\left(s\right)$$

$$=\sum_{j=1}^{J}\left[u_j\ln\frac{\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}+\left(1-u_j\right)\ln\frac{1}{1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)}\right.$$

$$+\ln\left(2\pi\right)^{-\frac{1}{2}}+\ln\left(\sigma\right)^{-1}+\ln\left(t_j\right)^{-1}+\ln\exp\left(-\frac{\left[\ln t_j-\left(\upsilon+s+r_j+g\left(1.7a_j\left(\theta-b_j\right)\right)\right)\right]^2}{2\sigma^2}\right)$$

$$\left.+\ln\left(2\pi\right)^{-\frac{1}{2}}+\ln\exp-\left(\frac{\theta^2}{2}\right)+\ln\left(2\pi\right)^{-\frac{1}{2}}+\ln\left(\sigma_s\right)^{-1}+\ln\exp-\left(\frac{s^2}{2\sigma_s^2}\right)\right]$$

$$=\sum_{j=1}^{J}\left[\left(u_j\left(\ln\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)-\ln\left[1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)\right]\right)\right)\right.$$

$$+\left(1-u_j\right)\left(\ln 1-\ln\left[1+\exp\left(1.7a_j\left(\theta-\eta t_j-b_j\right)\right)\right]\right)$$

$$-\frac{1}{2}\ln 2\pi-\ln\sigma-\ln t_j+\left(-\frac{\left[\ln t_j-\left(\upsilon+s+r_j+g\left(1.7a_j\left(\theta-b_j\right)\right)\right)\right]^2}{2\sigma^2}\right)$$

$$\left.-\frac{1}{2}\ln 2\pi-\left(\frac{\theta^2}{2}\right)-\frac{1}{2}\ln 2\pi-\ln\sigma_s-\left(\frac{s^2}{2\sigma_s^2}\right)\right].$$

(3.37)

$p$ now denotes the log of the posterior. Taking the derivative according to $\theta$, therefore it follows that

$$\begin{aligned}
\frac{\partial p}{\partial \theta} &= \sum_{j=1}^{J} \left\{ \left(u_j\right) \left[ \left(1.7a_j\right) - \frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta - \eta t_j - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta - \eta t_j - b_j\right)\right)} \right] \right. \\
&\quad + \left(1-u_j\right)\left[ - \frac{\left(1.7a_j\right)\exp\left(1.7a_j\left(\theta - \eta t_j - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta - \eta t_j - b_j\right)\right)} \right] \\
&\quad \left. \left( \frac{2\left[\ln t_j - \left(\upsilon + s + r_j + g\left(1.7a_j\left(\theta - b_j\right)\right)\right)\right]\left(g1.7a_j\right)}{2\sigma^2} \right) - \theta \right\} \\
&= \sum_{j=1}^{J} \left[ \left(1.7a_j\right)u_{kj} - \frac{\left(1.7a_i\right)\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} \right. \\
&\quad \left. \frac{\left(g1.7a_j\right)\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]}{\sigma^2} \right] - \theta \\
&= \sum_{j=1}^{J} \left(1.7a_j\right) \left( u_{kj} - \frac{\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)}{1+\exp\left(1.7a_j\left(\theta_k - \eta t_{kj} - b_j\right)\right)} + \frac{g\left[\ln t_{kj} - \left(\upsilon + s_k + r_j + g\left(1.7a_j\left(\theta_k - b_j\right)\right)\right)\right]}{\sigma^2} \right) - \theta.
\end{aligned} \tag{3.38}$$

Taking the derivative according to $s$, therefore we obtain

$$\frac{\partial p}{\partial s} = \left( \sum_{j=1}^{J} \frac{\ln t_j - \left(\upsilon + s + r_j + g\left(1.7a_j\left(\theta - b_j\right)\right)\right)}{\sigma^2} \right) - \frac{s}{\sigma_s^2} \ . \tag{3.39}$$

The ability parameters $\left(\theta\right)$ and person slowness parameters $\left(s\right)$ are maximized by using the general purpose optimization algorithm, *BFGS* (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) implemented in *optim* R function.

## Simulation

Simulation studies were conducted to evaluate the precision of item and person parameter estimates of the proposed extended item response theory model incorporating item response time. Three different simulation designs are taken into consideration for the model presented:  item parameter estimation (Equations 3.27 – 3.35) in simulation design 1, person parameter estimation (Equations 3.36 – 3.39) in simulation design 2, and person parameter estimation based on true and estimated item parameters (Equations 3.27 – 3.39) in simulation design 3. The simulations were performed using an R program which calls a Fortran program.

### Simulation design 1 – item parameter estimation using MML

The main interest of this simulation is to determine how well the marginal maximum likelihood (MML) method can estimate the item parameters when the response accuracy and response time data are generated from assumed model with discrete distributions of ability and person slowness parameters.  Equations 3.27 – 3.35 are employed.

In the data generation phase, the ability parameter $\theta$ and the person slowness parameter $s$ are two latent variables which are taken to be discrete. The model is

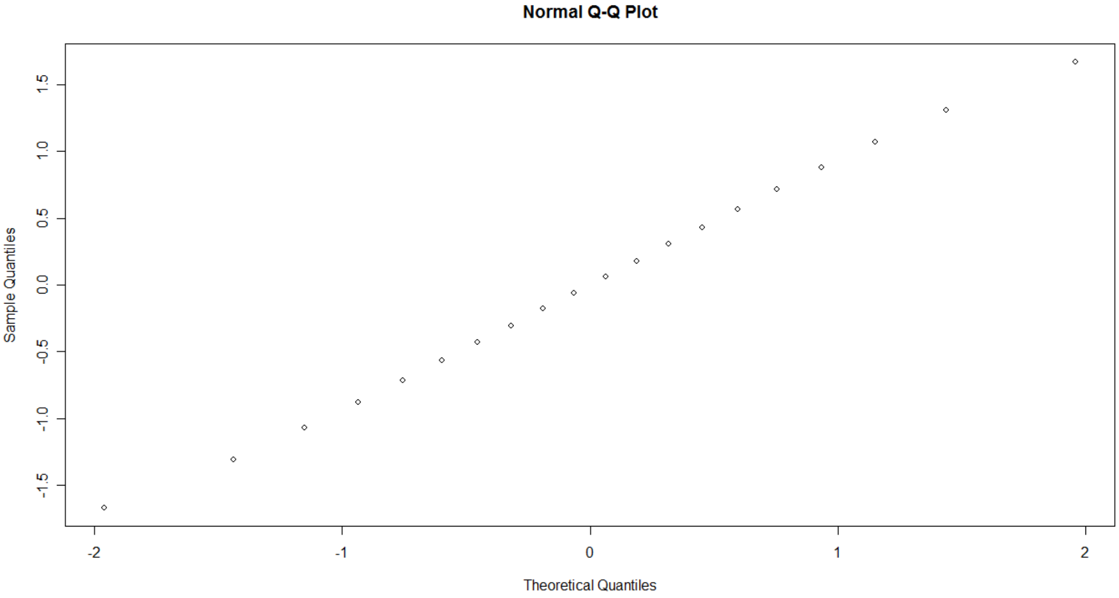


Figure 3.3 The normal probability plot values of $q_1,\ldots,q_{20}$ and $s_1,\ldots,s_{20}$

specified as follows: The ability parameter $\theta \in \{\sigma_q q_1, \sigma_q q_2, \ldots, \sigma_q q_K\}$ is assumed to be from discrete $N(0, \sigma_q)$ where $\sigma_q$ is the known scale factor and fixed to one. The values $q_1, \ldots, q_k$ are chosen to resemble normal distribution $N(0,1)$. The person slowness parameter $s \in \{\sigma_s s_1, \sigma_s s_2, \ldots, \sigma_s s_L\}$ is assumed to be from discrete $N(0, \sigma_s)$ where $\sigma_s$ is an unknown scale factor. The values $s_1, \ldots, s_L$ are known and to resemble $N(0,1)$. Namely, the model approximated the normal distribution for ability parameter $\theta$ by $K =$ 20 values with equal probability of $1/20$ with spacing chosen to resemble to normal distribution. The same is done for slowness parameter $s$ using $L$ = 20 values with equal probability of $1/20$. The Figure 3.3 displays the normal probability plot of the values of $q_1, \ldots, q_{20}$ and $s_1, \ldots, s_{20}$ showing that these 20 values may be regarded as forming an idealized normal sample. The 20 values forming a discrete approximation to a normal

Table 3.1 20 values forming a discrete approximation to a normal distribution which is used for both ability and person slowness parameters

| Values |
|---|
| -1.6683912 |
| -1.3091717 |
| -1.0675705 |
| -0.8761428 |
| -0.7124430 |
| -0.5659488 |
| -0.4307273 |
| -0.3029804 |
| -0.1800124 |
| -0.0597171 |
| 0.0597171 |
| 0.1800124 |
| 0.3029804 |
| 0.4307273 |
| 0.5659488 |
| 0.7124430 |
| 0.8761428 |
| 1.0675705 |
| 1.3091717 |
| 1.6683912 |

distribution which is used for both person parameters are shown in the Table 3.1. The 20 levels of the latent variables are used because the research indicates that 20 levels of the latent variable are about the minimum number needed in obtaining a reasonable estimate for the continuous normal distribution (Titterington, Smith & Markov, 1985). It is assumed that $(\theta, s)$ are independently distributed in the simulated data (See Table D.1).

The same set of true values of item parameters are preset to generate the response pattern and response time for 20 items and 40 items respectively (See Table 3.2 & Table 3.3). The true item discrimination parameters $(a)$ are fixed for 20 or 40 items and within range of 0.6453 to 1.78. The true item difficulty parameters $(b)$ are fixed for 20 or 40 items, lower bound -2 and upper bound 2. The item slowness parameters $(r)$ are fixed and ranging from 0.5019 to 1.0 for 20 items or 40 items. The values of item slowness parameter$(r)$ are chosen based on the similar values chosen by Wang and Hanson (2006).

Table 3.2 True Item Parameters ( $a, b, r$ ) used in Simulation for 20 items

| Item | $a$ | $b$ | $r$ |
|---|---|---|---|
| 1 | 1.0036 | -2.0000 | 0.7055 |
| 2 | 1.0650 | -1.7895 | 0.5795 |
| 3 | 0.9497 | -1.5789 | 0.5019 |
| 4 | 1.2234 | -1.3684 | 1.0000 |
| 5 | 0.9407 | -1.1579 | 0.8145 |
| 6 | 0.6576 | -0.9474 | 0.5445 |
| 7 | 0.8279 | -0.7368 | 0.8626 |
| 8 | 0.9844 | -0.5263 | 0.7350 |
| 9 | 1.5805 | -0.3158 | 0.8714 |
| 10 | 1.2155 | -0.1053 | 0.9496 |
| 11 | 0.7860 | 0.1053 | 0.5249 |
| 12 | 1.3716 | 0.3158 | 0.5350 |
| 13 | 1.2027 | 0.5263 | 0.6874 |
| 14 | 1.5569 | 0.7368 | 0.6227 |
| 15 | 1.2431 | 0.9474 | 0.8326 |
| 16 | 0.6453 | 1.1579 | 0.8298 |
| 17 | 1.4352 | 1.3684 | 0.7892 |
| 18 | 0.8156 | 1.5789 | 0.9110 |
| 19 | 1.7798 | 1.7895 | 0.6951 |
| 20 | 0.9110 | 2.0000 | 0.5439 |

Table 3.3 True Item Parameters ( $a, b, r$ ) used in Simulation for 40 items

| Item | $a$ | $b$ | $r$ |
|---|---|---|---|
| 1 | 1.0036 | -2.0000 | 0.7055 |
| 2 | 1.0320 | -1.8974 | 0.6760 |
| 3 | 1.0650 | -1.7949 | 0.5795 |
| 4 | 0.9860 | -1.6923 | 0.5340 |
| 5 | 0.9497 | -1.5897 | 0.5019 |
| 6 | 1.1230 | -1.4872 | 0.8980 |
| 7 | 1.2234 | -1.3846 | 1.0000 |
| 8 | 0.9670 | -1.2821 | 0.9230 |
| 9 | 0.9407 | -1.1795 | 0.8145 |
| 10 | 0.8760 | -1.0769 | 0.7860 |
| 11 | 0.6576 | -0.9744 | 0.5445 |
| 12 | 0.7560 | -0.8718 | 0.6780 |
| 13 | 0.8279 | -0.7692 | 0.8626 |
| 14 | 0.9340 | -0.6667 | 0.7920 |
| 15 | 0.9844 | -0.5641 | 0.7350 |
| 16 | 1.2450 | -0.4615 | 0.8120 |
| 17 | 1.5805 | -0.3590 | 0.8714 |
| 18 | 1.4350 | -0.2564 | 0.9230 |
| 19 | 1.2155 | -0.1538 | 0.9496 |
| 20 | 0.9870 | -0.0513 | 0.6780 |
| 21 | 0.7860 | 0.0513 | 0.5249 |
| 22 | 1.1240 | 0.1538 | 0.5950 |
| 23 | 1.3716 | 0.2564 | 0.5350 |
| 24 | 1.1000 | 0.3590 | 0.6230 |
| 25 | 1.2027 | 0.4615 | 0.6874 |
| 26 | 1.3450 | 0.5641 | 0.6920 |
| 27 | 1.5569 | 0.6667 | 0.6227 |
| 28 | 1.4320 | 0.7692 | 0.7970 |
| 29 | 1.2431 | 0.8718 | 0.8326 |
| 30 | 0.8780 | 0.9744 | 0.8010 |
| 31 | 0.6453 | 1.0769 | 0.8298 |
| 32 | 1.2220 | 1.1795 | 0.7230 |
| 33 | 1.4352 | 1.2821 | 0.7892 |
| 34 | 0.7870 | 1.3846 | 0.8930 |
| 35 | 0.8156 | 1.4872 | 0.9110 |
| 36 | 1.5670 | 1.5897 | 0.7520 |
| 37 | 1.7798 | 1.6923 | 0.6951 |
| 38 | 1.4560 | 1.7949 | 0.5930 |
| 39 | 0.9110 | 1.8974 | 0.5439 |
| 40 | 0.7980 | 2.0000 | 0.6250 |

The additional parameters estimated in the simulation are $\eta, g, \sigma,$ and $\sigma_s$. True values of these parameters are shown in Table 3.4. Using the person parameters, item parameters and the additional parameters, response time data are generated under the assumption that the proposed lognormal marginal distribution (Equation 3.16) is the correct model. Item response data are then generated according to the following method: first, the probability of a correct response to item $j$ by examinee $i$ given the response time $t_{ij}$, $P_{ij}$, is calculated based on the proposed conditional distribution (Equation 3.12). Then, a random number $p$ from the uniform distribution $U(0,1)$ is independently generated. If $P_{ij} > p$, then a correct response is obtained for the item $j$ by examinee $i$ given the response time $t_{ij}$, otherwise an incorrect response is obtained. These simulated response data and response time data set are the target data set in the simulation studies.

The test length in this simulation is set to be either 20 or 40 items. The number of simulated examinees is 1000 and 2000 in this study. Given these factors, four combinations are specified. 100 replications are performed for each of four simulated condition.

The marginal maximum likelihood is computed and maximized directly using R program. In R program, *BFGS* method is implemented as one of the option inputs to the *optim* function (R Development of Core Team, 2007).

**Simulation design 2 – person parameter estimation using MAP**

The goal of this simulation is to test out how well the maximum a posteriori (MAP) procedure behaves in estimating the continuous ability and person slowness parameters when the item parameters are known. Equations 3.36 – 3.39 are utilized.

Table 3.4 Additional parameters ($\eta, g, \sigma, \sigma_s$) used to generate the data

| Parameter | True Value |
|---|---|
| $\eta$ | 0.002 |
| $g$ | 0.5 |
| $\sigma$ | 1 |
| $\sigma_s$ | 1 |

Unlike simulation design 1, the continuous distribution is assumed for both ability and person slowness parameters to generate the data because in practice these person parameters of a randomly selected examinee are considered to be continuous. To check the effect of sampling from the whole range of the ability and person slowness parameters, the person parameter recoveries are investigated in the simulation studies. The same set of true values of item parameters and additional parameters are preset to generate the response pattern and response time for 20 items and 40 items respectively (See Table 3.2, Table 3.3 & Table 3.4). The technique of generating the data for response time and response pattern is equivalent to the approach explained in the simulation design 1, but in this study the ability and person slowness are assumed to come from continuous distributions.

The MAP procedure is implemented to estimate the values of ability parameters $(\theta)$ and person slowness parameters $(s)$ that maximize the posterior distribution. The prior distribution assumes that $\theta$ and $s$ are independent with $\theta \sim N(0,1)$ and $s \sim N(0,\sigma_s^2)$ respectively, where $\sigma_s$ assumes to be a unit. The ability parameters $(\theta)$ and person slowness parameters $(s)$ are maximized by using the general purpose optimization algorithm, *BFGS* (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) implemented in *optim* R function.

In this simulation, the test length is set to be either 20 or 40 items. The number of simulated examinees is 1000 and 2000. Thus, four combinations are specified. 100 replications are performed for each of four simulated condition.

**Simulation design 3 – person parameter recovery using MML and MAP**

This simulation intends to see how well ability and person slowness parameters are estimated using MAP procedure based on two scenarios, either when the true item parameters are known or when item parameters are estimated. Equations 3.27 – 3.39 are employed.

In the data generation phase, the ability parameter $\theta$ and the person slowness parameter $s$ are assumed to be discrete. The condition of sampling from discrete for person parameters are the same as stated in the simulation design 1. Discretizing the distributions enabled the model to be implemented in the straightforward manner. In other words, sampling the data from discrete and fitting the model to the discrete lead to the finite sum and easy of integration for marginal maximum likelihood method which enables the ability parameter to be estimated easier and faster than when the continuous distribution is assumed. The set of true values of item parameters and additional parameters introduced in Table 3.2 and Table 3.4, respectively, are employed to generate the response pattern and response time. The way of generating the data for response time and response pattern is the same as described in the simulation design 1.

In the estimation phase, both true and estimated item parameters are used to estimate person parameters. To be precise, in the first scenario where the known true item parameters are used to estimate person parameters, the MAP procedure (Equations 3.36 – 3.39) is implemented to estimate person parameters. In the second scenario where the estimated item parameters are used to estimate person parameters, two stages are involved in the estimation phase: First, the MML method (Equations 3.27 – 3.35) is implemented to estimate the item parameters. To estimate the item parameters, the marginal maximum likelihood is computed and maximized directly. Then, using the estimated item parameters, the MAP procedure (Equations 3.36 – 3.39) is applied to estimate person parameters. The ability parameters $(\theta)$ and person slowness parameters $(s)$ are maximized by using the general purpose optimization algorithm, BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) implemented in R function optim.

One simulation condition – a 20 item test for 1000 examinees – is studied. The process is repeated 100 times for each scenario.

**Criteria for comparisons**

The simulation studies are evaluated according to the following comparison criteria: mean and standard deviation of the RMSE, bias, and standard error (SE), sample variance of the item parameter estimates and mean of the variance estimates based on the Fisher information and corresponding standard deviation**,** and means and standard deviations of the correlations.

In the simulation design 1, the average of bias, average of standard error (SE) of estimates and average of root mean square error (RMSE) are computed for each item parameter and additional parameter across 100 replications**.** The root mean squared error

(RMSE) of the estimated parameters is commonly used as a criterion for the recovery of item parameters in simulation studies. The RMSE indicates the total error in the parameter recovery. It is composed of the squared bias and squared standard error of estimates (See Equation 3.40)

$$\left(RMSE\right)^2 = \left(bias\right)^2 + \left(SE\right)^2 . \tag{3.40}$$

The RMSE is the square root of the average of the squared deviations of estimated parameters from the corresponding true values. Let $\psi_j$ represent a parameter of item $j$, i.e. $a_j, b_j,$ or $r_j$, and $\hat{\psi}_{jm}$ be the estimate of $\psi_j$ from the $m^{\text{th}}$ replication for $j = 1, \dots, J$ and $m = 1, \dots, M$ . Here $J$ denotes the number of items and $M$ stands for the number of replications where $M = 100$ in this simulation. For each item parameter, RMSE is defined as

$$RMSE\left(\hat{\psi}_j\right) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\psi}_{jm} - \psi_j\right)^2} . \tag{3.41}$$

The bias is the difference between the mean estimated item parameter values $\hat{\psi}_{jm}$ across $M = 100$ replications and the true item parameter value $\psi_j$ for a particular item parameter. It is given as

$$Bias\left(\hat{\psi}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_{jm} - \psi_j . \tag{3.42}$$

The SE is the standard deviation of the estimated parameter values across $M = 100$ replications, expressed as

$$SE\left(\hat{\psi}_j\right) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\psi}_{jm} - \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_{jm}\right)^2} . \tag{3.43}$$

In addition**,** the sample variance of the item parameter estimates and the mean of the variance estimates based on the Fisher information and corresponding standard deviation are compared to check the consistency of the estimates.

In simulation designs 2 and 3, the means of the true and estimated person parameters and their variability with corresponding mean of standard deviations and their variability from 100 replications are computed. While, four simulation conditions are studied in the second simulation, one simulation condition is studied in the third simulation. In addition, the correlations between true and estimated abilities and person slowness parameters are calculated for each parameter for each replication. Means and standard deviations of the correlations across replications are computed. Histograms are used to illustrate the distributions of the correlations.

# CHAPTER 4

# RESULTS

The results from the three simulation designs are discussed in this chapter. First, I present the results of the simulation design 1 – item parameter estimation, next, the results of the person parameter estimation in simulation design 2, followed by the results of the person parameter estimation based on true and estimated item parameters in simulation design 3. The quality of parameter estimates is investigated across various simulation conditions based on 100 replications.

Table 4.1 Across-item bias, SE and RMSE of the item discrimination parameter $(a)$ average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimate | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 1.0036 | 1.0066 | 0.0735 | 0.0030 | 0.0732 |
| 2 | 1.0650 | 1.0754 | 0.0765 | 0.0104 | 0.0768 |
| 3 | 0.9497 | 0.9489 | 0.0674 | -0.0008 | 0.0670 |
| 4 | 1.2234 | 1.2232 | 0.0747 | -0.0002 | 0.0743 |
| 5 | 0.9407 | 0.9425 | 0.0771 | 0.0018 | 0.0767 |
| 6 | 0.6576 | 0.6656 | 0.0687 | 0.0080 | 0.0689 |
| 7 | 0.8279 | 0.8311 | 0.0702 | 0.0032 | 0.0699 |
| 8 | 0.9844 | 0.9962 | 0.0692 | 0.0118 | 0.0698 |
| 9 | 1.5805 | 1.5867 | 0.0870 | 0.0062 | 0.0868 |
| 10 | 1.2155 | 1.2204 | 0.0695 | 0.0049 | 0.0693 |
| 11 | 0.7860 | 0.8050 | 0.0658 | 0.0190 | 0.0682 |
| 12 | 1.3716 | 1.3703 | 0.0816 | -0.0013 | 0.0812 |
| 13 | 1.2027 | 1.2079 | 0.0684 | 0.0052 | 0.0682 |
| 14 | 1.5569 | 1.5494 | 0.0783 | -0.0075 | 0.0782 |
| 15 | 1.2431 | 1.2414 | 0.0730 | -0.0017 | 0.0727 |
| 16 | 0.6453 | 0.6462 | 0.0608 | 0.0009 | 0.0605 |
| 17 | 1.4352 | 1.4296 | 0.0803 | -0.0056 | 0.0801 |
| 18 | 0.8156 | 0.8204 | 0.0696 | 0.0048 | 0.0695 |
| 19 | 1.7798 | 1.7804 | 0.0954 | 0.0006 | 0.0949 |
| 20 | 0.9110 | 0.9134 | 0.0694 | 0.0024 | 0.0691 |

### Simulation design 1 – item parameter estimation using MML

Table 4.1 results show bias, SE and RMSE of the item discrimination parameter ($a$) average estimates based on 100 replications for 20 items and 1000 examinees. The estimates contain the results of the average of the estimates from 100 replications. Comparing the values of true and estimated item discrimination parameters reveal that the estimates are fairly close to their corresponding true values. This is confirmed in the small biases for the item discrimination parameter. The RMSE of the estimated item discrimination parameters, which is the square root of the average of the squared deviations of estimated item discriminations from their corresponding true item discriminations based on 100 replications, are small values. The standard deviations of the 100 estimates (SE) are small across 20 items. In general, the average bias, the average SE and the average RMSE tend to decrease as sample size increases from

Table 4.2 Sample variance of the item discrimination ($a$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_a$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_a$ (SD) |
|---|---|---|
| 1 | 0.005401 | 0.005730 (0.000455) |
| 2 | 0.005850 | 0.005792 (0.000487) |
| 3 | 0.004539 | 0.005188 (0.000367) |
| 4 | 0.005583 | 0.006013 (0.000499) |
| 5 | 0.005938 | 0.004892 (0.000336) |
| 6 | 0.004726 | 0.004220 (0.000249) |
| 7 | 0.004922 | 0.004453 (0.000287) |
| 8 | 0.004784 | 0.004759 (0.000339) |
| 9 | 0.007569 | 0.007086 (0.000640) |
| 10 | 0.004825 | 0.005349 (0.000411) |
| 11 | 0.004335 | 0.004267 (0.000255) |
| 12 | 0.006666 | 0.005962 (0.000483) |
| 13 | 0.004676 | 0.005376 (0.000428) |
| 14 | 0.006126 | 0.007023 (0.000654) |
| 15 | 0.005333 | 0.005733 (0.000470) |
| 16 | 0.003702 | 0.004258 (0.000259) |
| 17 | 0.006445 | 0.006916 (0.000644) |
| 18 | 0.004851 | 0.004819 (0.000317) |
| 19 | 0.009099 | 0.009778 (0.001107) |
| 20 | 0.004823 | 0.005375 (0.000390) |

1000 to 2000 examinees and number of items increases from 20 to 40 items (See Tables A.1, A.9 & A.17).

In addition, a consistency check is provided in Table 4.2 where the means of the estimated variances based on the Fisher information ($\overline{I}_a$) are pretty close to the corresponding sample variances of the item discrimination estimates ($S^2$) over 100 simulation replications for 20 items and 1000 examinees. The variability of the variance estimates presented as standard deviations (SD) is small. Therefore, it is concluded that the item discrimination parameter ($a$) average estimates based on 100 replications for 20 items and 1000 examinees are consistent and pretty accurate. In addition, the variance and the variability of the variance exhibited in Tables A.2, A.10 and A.18 decrease as sample size increases from 1000 to 2000 examinees and the number of items increases from 20 to 40 items.

Table 4.3 Across-item bias, SE and RMSE of the item difficulty parameter ($b$) average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | -2.0000 | -2.0005 | 0.1519 | -0.0005 | 0.1512 |
| 2 | -1.7895 | -1.7833 | 0.1334 | 0.0062 | 0.1328 |
| 3 | -1.5789 | -1.5893 | 0.1254 | -0.0104 | 0.1252 |
| 4 | -1.3684 | -1.3760 | 0.0904 | -0.0075 | 0.0902 |
| 5 | -1.1579 | -1.1532 | 0.1088 | 0.0047 | 0.1084 |
| 6 | -0.9474 | -0.9454 | 0.1313 | 0.0020 | 0.1307 |
| 7 | -0.7368 | -0.7346 | 0.1037 | 0.0022 | 0.1032 |
| 8 | -0.5263 | -0.5218 | 0.0760 | 0.0045 | 0.0757 |
| 9 | -0.3158 | -0.3025 | 0.0554 | 0.0132 | 0.0567 |
| 10 | -0.1053 | -0.0960 | 0.0604 | 0.0093 | 0.0608 |
| 11 | 0.1053 | 0.1200 | 0.0841 | 0.0147 | 0.0850 |
| 12 | 0.3158 | 0.3114 | 0.0622 | -0.0044 | 0.0620 |
| 13 | 0.5263 | 0.5300 | 0.0690 | 0.0036 | 0.0688 |
| 14 | 0.7368 | 0.7415 | 0.0618 | 0.0046 | 0.0617 |
| 15 | 0.9474 | 0.9616 | 0.0797 | 0.0142 | 0.0805 |
| 16 | 1.1579 | 1.1682 | 0.1290 | 0.0103 | 0.1287 |
| 17 | 1.3684 | 1.3740 | 0.0899 | 0.0056 | 0.0896 |
| 18 | 1.5789 | 1.5889 | 0.1245 | 0.0099 | 0.1242 |
| 19 | 1.7895 | 1.8208 | 0.1020 | 0.0313 | 0.1062 |
| 20 | 2.0000 | 2.0140 | 0.1769 | 0.0140 | 0.1766 |

The tendencies shown in Tables 4.1 and 4.2 are demonstrated similarly in three other conditions, i.e. in Tables A.1 and A.2 for 20 items and 2000 examinees, in Tables A.9 and A.10 for 40 items and 1000 examinees, and in Tables A.17 and A.18 for 40 items and 2000 examinees. Across the four simulation conditions, it is noted that the effect of number of examinees on the item discrimination parameter ($a$) average estimates is larger than the effect of test length.

Across-item bias, SE and RMSE of the item difficulty parameter ($b$) average estimates based on 100 replications for 20 items and 1000 examinees are shown in Table 4.3. Comparing the values of true and estimated item difficulty parameters uncover that the estimates are reasonably close to their corresponding true values. The obtained average bias, average SE and average RMSE are fairly small. Similar findings are shown

Table 4.4 Sample variance of the item difficulty ($b$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_b$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_b$ (SD) |
|---|---|---|
| 1 | 0.012071 | 0.011872 (0.002317) |
| 2 | 0.009504 | 0.008680 (0.001477) |
| 3 | 0.008578 | 0.008725 (0.001501) |
| 4 | 0.004076 | 0.004584 (0.000457) |
| 5 | 0.005412 | 0.005888 (0.000840) |
| 6 | 0.008556 | 0.009269 (0.001685) |
| 7 | 0.005469 | 0.005202 (0.000769) |
| 8 | 0.003629 | 0.003336 (0.000320) |
| 9 | 0.001681 | 0.001615 (0.000073) |
| 10 | 0.001927 | 0.002129 (0.000128) |
| 11 | 0.003676 | 0.004096 (0.000436) |
| 12 | 0.001676 | 0.001891 (0.000091) |
| 13 | 0.002757 | 0.002449 (0.000193) |
| 14 | 0.002071 | 0.002012 (0.000124) |
| 15 | 0.002934 | 0.003148 (0.000313) |
| 16 | 0.013395 | 0.011765 (0.002877) |
| 17 | 0.003683 | 0.003730 (0.000383) |
| 18 | 0.012292 | 0.011272 (0.002327) |
| 19 | 0.004014 | 0.004510 (0.000450) |
| 20 | 0.014045 | 0.013638 (0.002906) |

in three other simulation conditions in Tables A.3, A.11 and A.19. Comparing results from the three other simulation conditions displayed in Tables A.3, A.11 and A.19, the bias, SE and RMSE tend to decrease as sample size and number of items increase.

The consistency is shown in Table 4.4 where the means of the estimated variances based on the Fisher information ($\overline{I}_b$) are pretty close to the corresponding sample variances of the item difficulty ($b$) estimates over 100 simulation replications for 20 items and 1000 examinees. Also, the variability of the variance estimates presented in standard deviations (SD) is very small. As a result, this indicates that average item difficulty ($b$) estimates based on 100 replications for 20 items and 1000 examinees are pretty precise. Typically, the greater the number of examinees and the longer the test length, the better the estimated item difficulty parameters ($b$) are. However, increasing the number of examinees is a more effective way to produce better item difficulty parameters ($b$) average estimates than increasing the test length.

Table 4.5 Across-item bias, SE and RMSE of the item slowness parameter ($r$) average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 0.7055 | 0.6976 | 0.0899 | -0.0079 | 0.0898 |
| 2 | 0.5795 | 0.5667 | 0.0881 | -0.0128 | 0.0886 |
| 3 | 0.5019 | 0.4936 | 0.0740 | -0.0083 | 0.0741 |
| 4 | 1.0000 | 0.9890 | 0.0744 | -0.0110 | 0.0748 |
| 5 | 0.8145 | 0.8121 | 0.0640 | -0.0024 | 0.0637 |
| 6 | 0.5445 | 0.5383 | 0.0590 | -0.0062 | 0.0590 |
| 7 | 0.8626 | 0.8595 | 0.0543 | -0.0031 | 0.0542 |
| 8 | 0.7350 | 0.7315 | 0.0606 | -0.0035 | 0.0604 |
| 9 | 0.8714 | 0.8762 | 0.0593 | 0.0048 | 0.0592 |
| 10 | 0.9496 | 0.9519 | 0.0472 | 0.0023 | 0.0471 |
| 11 | 0.5249 | 0.5295 | 0.0531 | 0.0046 | 0.0530 |
| 12 | 0.5350 | 0.5298 | 0.0547 | -0.0052 | 0.0547 |
| 13 | 0.6874 | 0.6920 | 0.0569 | 0.0046 | 0.0568 |
| 14 | 0.6227 | 0.6246 | 0.0608 | 0.0019 | 0.0605 |
| 15 | 0.8326 | 0.8412 | 0.0588 | 0.0086 | 0.0591 |
| 16 | 0.8298 | 0.8389 | 0.0617 | 0.0091 | 0.0620 |
| 17 | 0.7892 | 0.7947 | 0.0757 | 0.0055 | 0.0755 |
| 18 | 0.9110 | 0.9211 | 0.0625 | 0.0101 | 0.0630 |
| 19 | 0.6951 | 0.7360 | 0.1000 | 0.0409 | 0.1076 |
| 20 | 0.5439 | 0.5591 | 0.0882 | 0.0152 | 0.0890 |

Results similar to those shown in Tables 4.3 and 4.4 are obtained in three other simulation conditions, i.e. in Tables A.3 and A.4 for 20 items and 2000 examinees, in Tables A.11 and A.12 for 40 items and 1000 examinees, and in Tables A.19 and A.20 for 40 items and 2000 examinees.

Across-item bias, SE and RMSE of the item slowness parameter $(r)$ average estimates based on 100 replications for 20 items and 1000 examinees are given in Table 4.5. The item slowness parameter $(r)$ average estimates are pretty close to their corresponding true values. This is supported by the small average bias, SE and RMSE. Similar findings of the item slowness parameter $(r)$ are given in three other simulation conditions (See Tables A.5, A.13 & A.21). Across all four simulation conditions, the bias, SE and RMSE are likely to decrease as sample size as well as number of items increase (See Tables 4.5, A.5, A.13 and A.21).

Table 4.6 Sample variance of the item slowness $(r)$ estimates $(S^2)$ and mean of the variance estimates based on the Fisher information $(\overline{I}_r)$ and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_r$ (SD) |
|---|---|---|
| 1 | 0.003560 | 0.003729 (0.000271) |
| 2 | 0.003676 | 0.003464 (0.000245) |
| 3 | 0.002778 | 0.002667 (0.000174) |
| 4 | 0.002964 | 0.002961 (0.000201) |
| 5 | 0.001978 | 0.002084 (0.000115) |
| 6 | 0.001469 | 0.001577 (0.000071) |
| 7 | 0.001521 | 0.001602 (0.000076) |
| 8 | 0.001979 | 0.001560 (0.000071) |
| 9 | 0.001934 | 0.001648 (0.000069) |
| 10 | 0.001515 | 0.001460 (0.000055) |
| 11 | 0.001136 | 0.001375 (0.000049) |
| 12 | 0.001234 | 0.001561 (0.000064) |
| 13 | 0.001387 | 0.001647 (0.000074) |
| 14 | 0.001956 | 0.002092 (0.000111) |
| 15 | 0.002519 | 0.002149 (0.000114) |
| 16 | 0.002075 | 0.001675 (0.000082) |
| 17 | 0.002903 | 0.003334 (0.000193) |
| 18 | 0.002512 | 0.002330 (0.000142) |
| 19 | 0.006648 | 0.005921 (0.000422) |
| 20 | 0.002761 | 0.003277 (0.000213) |

The consistency is assessed in Table 4.6 where the means of the estimated variances based on the Fisher information are pretty close to their corresponding sample variances of the item slowness parameter ($r$) estimates. The obtained standard deviation is very small, providing support that item slowness parameter ($r$) average estimates based on 100 replications for 20 items and 1000 examinees are pretty precise. Similar trends in Tables 4.6 are shown in three other simulation conditions (See Tables A.6, A.14 & A.22). The greater the number of examinees and the longer the test length, the better the estimated item slowness parameters ($r$) are. The influence of number of examinees on the item slowness parameter ($r$) average estimates is larger than the influence of test length.

Across-item bias, SE and RMSE of the parameter ($\eta, g, \sigma, \sigma_s$) average estimates based on 100 replications for 20 items and 1000 examinees are given in Table 4.7. The parameter ($\eta, g, \sigma, \sigma_s$) average estimates are pretty close to their corresponding true values. This is supported by the small average bias, average SE and average RMSE. Similar findings are reported in three other simulation conditions (See Tables A.7, A.15 & A.23). Given Tables 4.7, A.7, A.15 and A.23, the bias, SE and RMSE are likely to decrease as sample size and number of items increase.

Table 4.8 shows that the means of the estimated variances based on the Fisher information are pretty close to the corresponding sample variances of the $\eta, g, \sigma, \sigma_s$ parameter estimates for 20 items and 1000 examinees. The SD is small. It gives support that $\eta, g, \sigma, \sigma_s$ parameter average estimates based on 100 replications for 20 items and 1000 examinees are consistent and accurate. Similar tendencies are found in the three other simulation conditions (See Tables A.8, A.16 & A.24).

Table 4.7 Across-item bias, SE and RMSE of the parameter ($\eta, g, \sigma, \sigma_s$) average estimates based on 100 replications for 20 items and 1000 examinees

| Parameter | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| $\eta$ | 0.002 | 0.0020 | 0.0010 | -0.00003 | 0.0010 |
| $g$ | 0.5 | 0.5050 | 0.0284 | 0.0050 | 0.0287 |
| $\sigma$ | 1 | 0.9995 | 0.0054 | -0.0005 | 0.0054 |
| $\sigma_s$ | 1 | 0.9954 | 0.0195 | -0.0046 | 0.0199 |

Table 4.8 Sample variance of the parameter ($\eta, g, \sigma, \sigma_s$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\bar{I}$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 1000 examinees

| Parameter | $S^2$ | $\bar{I}$ (SD) |
|---|---|---|
| $\eta$ | 0.00000051 | 0.00000056 (0.00000011) |
| $g$ | 0.00043301 | 0.00042891 (0.00003321) |
| $\sigma$ | 0.00001249 | 0.00001327 (0.00000009) |
| $\sigma_s$ | 0.00023050 | 0.00021386 (0.00000816) |

Generally, the greater the number of examinees and the longer the test length, the better the estimated parameter ($\eta, g, \sigma, \sigma_s$) are. However, the parameter ($\eta, g, \sigma, \sigma_s$) average estimates are affected more by the number of examinees than by the test length.

In summary, the marginal maximum likelihood (MML) method estimated the item parameters and additional parameters fairly well when the response accuracy and response time data were generated from the assumed model with discrete distributions of ability and person slowness parameters. The bias, SE and RMSE were reasonably small across all four simulation conditions. The observed variances of the estimates indicated that the parameter average estimates based on 100 replications were consistent and accurate across four simulation conditions. In general, the greater the number of examinees and the longer the test length, the better the parameter estimates. However, it was found that the impact of number of examinees on the parameter average estimates was larger than the impact of test length.

### Simulation design 2 – person parameter estimation using MAP

To determine how well the maximum a posteriori (MAP) procedure behaves in estimating the continuous ability and person slowness parameters when the true item parameters are known, firstly, various means and standard deviations of two person parameters from 100 replications were observed. In Table 4.9, Mean (SD) columns are obtained as followings: True or estimated mean person parameters $\left(\theta, \hat{\theta}, s, \hat{s}\right)$ based on 1000 or 2000 examinees are computed per iteration, this process repeats 100 times. Then, the mean and its variability (SD) based on 100 replications are reported. The SD(SD) is obtained in the same way. As a result, the means of the mean parameters $\left(\theta, \hat{\theta}, s, \hat{s}\right)$ and their variability with corresponding means of the mean standard deviations and their variability (SD(SD)) from 100 replications are reported for four simulation conditions.

Because both abilities and person slowness parameters are assumed to be from standard normal distributions, the obtained means and standard deviations of the two person parameters across all four simulations conditions are fairly close to the assumed values, i.e. means of zero and standard deviations of one. Based on small variability values, it is conclude that the obtained mean values are consistent and pretty accurate. As the number of examinees and the test length increase, the obtained values are closer to the true values. However, the test length does not influence the parameter average estimates as much as the number of examinees across all four simulation conditions.

Table 4.9 Mean of the mean true abilities ($\theta$) from 100 replications and its variability, mean of the estimated mean abilities ($\hat{\theta}$) from 100 replications and its variability, mean of the mean true person slowness ($s$) from 100 replications and its variability, and mean of the estimated person slowness ($\hat{s}$) from 100 replications and its variability (Mean (SD)). Mean of the standard deviations and its variability (SD (SD)) from 100 replications with corresponding parameters ($\theta, \hat{\theta}, s, \hat{s}$) across four simulation conditions

| Item | Examinee | Parameter | Mean (SD) | SD (SD) |
|---|---|---|---|---|
| 20 | 1000 | $\theta$ | 0.0020 (0.0307) | 1.0015 (0.0221) |
| | | $\hat{\theta}$ | 0.0064 (0.0273) | 0.8939 (0.0199) |
| | | $s$ | -0.0015 (0.0303) | 1.0016 (0.0232) |
| | | $\hat{s}$ | -0.0036 (0.0288) | 0.9516 (0.0223) |
| | 2000 | $\theta$ | 0.0015 (0.0240) | 1.0014 (0.0165) |
| | | $\hat{\theta}$ | 0.0068 (0.0205) | 0.8933 (0.0141) |
| | | $s$ | -0.0012 (0.0226) | 0.9996 (0.0151) |
| | | $\hat{s}$ | -0.0044 (0.0218) | 0.9513 (0.0135) |
| 40 | 1000 | $\theta$ | -0.0023 (0.0342) | 0.9992 (0.0197) |
| | | $\hat{\theta}$ | 0.0027 (0.0315) | 0.9370 (0.0191) |
| | | $s$ | -0.0040 (0.0281) | 0.9998 (0.0230) |
| | | $\hat{s}$ | -0.0068 (0.0261) | 0.9730 (0.0221) |
| | 2000 | $\theta$ | -0.0020 (0.0205) | 0.9998 (0.0163) |
| | | $\hat{\theta}$ | 0.0013 (0.0199) | 0.9369 (0.0147) |
| | | $s$ | 0.0028 (0.0205) | 1.0005 (0.0168) |
| | | $\hat{s}$ | 0.0010 (0.0202) | 0.9736 (0.0163) |

Table 4.10 Correlation and standard deviation (SD) between true and estimated ability parameter ( $\theta, \hat{\theta}$ ) and between true and estimated person slowness parameter ( $s, \hat{s}$ ) across 100 replications

| Item | Examinee | Correlation (SD)<br>$\theta, \hat{\theta}$ | Correlation (SD)<br>$s, \hat{s}$ |
|---|---|---|---|
| 20 | 1000 | 0.9085 (0.0053) | 0.9502 (0.0034) |
| | 2000 | 0.9083 (0.0038) | 0.9501 (0.0021) |
| 40 | 1000 | 0.9484 (0.0029) | 0.9728 (0.0020) |
| | 2000 | 0.9481 (0.0022) | 0.9727 (0.0013) |

In Table 4.10, the correlations between true and estimated ability and person slowness parameters are reported. The correlations are obtained as follows: For each replication, the correlations are calculated between true and estimated person parameters. The means and standard deviations of these correlations across replications are computed.

Across four simulation conditions, the mean correlations between true and estimated ability parameters ranged from 0.9083 to 0.9484 and the correlations between true and estimated person slowness parameters ranged from 0.9501 to 0.9728. Overall, the correlations between true and estimated person slowness parameters are larger than the correlations of ability parameters across all four simulation conditions. In addition, the standard deviations of correlations are reasonably small for both person parameters across four simulation conditions. As the number of examinees and the number of items increase, the standard deviations of correlations decrease in both person parameter cases. The distributions of the correlations for ability and person slowness parameters are depicted in the histograms in Figures B.1 and B.2. Mostly, bell-shaped histograms are observed and many of the observations are around the obtained mean of correlations reported in Table 4.10. The high values of correlations and histograms indicate that the correlations are pretty consistent. Overall, the standard deviations of correlations for person slowness parameters are smaller than the standard deviations of correlations obtained for ability parameters. This demonstrates that in MAP procedures the person slowness parameters are estimated better than the ability parameters.

In general, the longer the test length, the better the correlations. It was found that the impact of test length on the average correlation estimates was larger than the impact of number of examinees. The standard deviations of correlations got smaller as the number of examinees and the number of items increase. However, it was discovered that the impact of number of examinees on the standard deviations of correlations was larger than the impact of test length.

In summary, the maximum a posteriori (MAP) procedure successfully estimates the continuous ability and person slowness parameters when the true item parameters are known. This is confirmed by the high and consistent values of correlations with correspondingly small variability.

### Simulation design 3 – person parameter recovery using MML and MAP

Under the simulation condition with 20 items and 1000 examinees, two scenarios are considered to see how well ability and person slowness parameters are estimated using the MAP procedure: first, when the true item parameters are known, followed by when item parameters are estimated.

Across-item bias, SE and RMSE of the item discrimination ($a$), item difficulty ($b$) and item slowness ($r$) parameters based on 100 replications for 20 items and 1000 examinees are presented in Tables C.1, C.2 and C.3 respectively.

Comparing the values of true and estimated item parameters based on the MML procedure shows that in general the estimates are reasonably close to their corresponding true values. The obtained average bias, average SE and average RMSE are small across 20 items from all three item parameters displayed in Tables C.1, C.2 and C.3, respectively. The standard deviations of the extreme item difficulty ($b$) values seem to be slightly larger than those around in the middle. The results for the additional parameters shown in Table C.4 also reveal small bias, SE and RMSE. The parameter estimates presented in Tables C.1, C.2, C.3 and C.4 are used to obtain person parameter estimations in scenario 2.

Table 4.11 Mean of true ($\theta$) and estimated mean abilities ($\hat{\theta}$) from 100 replications and its variability and mean of the standard deviations and its variability (SD(SD)) from 100 replications for 20 items and 1000 examinees

| | True $\theta$<br>Mean (SD) | $\hat{\theta}$<br>Using True<br>Item Parameters | $\hat{\theta}$<br>Using Estimated<br>Item Parameters |
|---|---|---|---|
| Mean (SD) | -0.0034 (0.0280) | 0.0020 (0.0262) | 0.0026 (0.0061) |
| SD (SD) | 0.8663 (0.0164) | 0.8061 (0.0142) | 0.8093 (0.0037) |

Table 4.12 Mean of true ($s$) and estimated mean person slowness ($\hat{s}$) from 100 replications and its variability (Mean (SD)). Mean of the standard deviations and its variability (SD(SD)) from 100 replications for 20 items and 1000 examinees

| | True $s$<br>Mean (SD) | $\hat{s}$<br>Using True<br>Item Parameters | $\hat{s}$<br>Using Estimated<br>Item Parameters |
|---|---|---|---|
| Mean (SD) | 0.0057 (0.0256) | 0.0037 (0.0262) | -0.0036 (0.0092) |
| SD (SD) | 0.8693 (0.0152) | 0.8377 (0.0141) | 0.8383 (0.0179) |

In Tables 4.11 and 4.12 the means of the true and estimated mean ability and person slowness parameters respectively from 100 replications and their variability for 20 items and 1000 examinees are reported. In Table 4.11, comparing the estimated mean ability with corresponding true mean ability indicates that means are very close and the variability of the means are fairly small. The means of standard deviations are comparable and their variability is quite small. In Table 4.12, comparing the estimated mean person slowness with corresponding true mean person slowness indicates that means are very close and the variability of the means are somewhat small. The means of standard deviations and its variability are alike. The variability of standard deviations seem to be very small. It implies that the obtained mean values are consistent and pretty accurate.

In Table 4.13, correlations and standard deviations (SD) between true and estimated person parameters across 100 replications. Comparing the correlations of person parameters based on true item parameters and estimated item parameters, correlations of the person parameters using true item parameters are almost same as those using estimated item parameters. The correlation between true and estimated ability parameter ($\theta, \hat{\theta}$) across 100 replications using true item parameters is 0.8881 whereas using estimated item parameters is 0.8873. Also, the correlation between true and estimated person slowness parameter ($s, \hat{s}$) across 100 replications using true item parameters is 0.9384, meanwhile using estimated item parameters is 0.9380. However, standard deviations are almost identical in both scenarios, using true item parameters or using estimated item parameters.

Table 4.13 Correlation and standard deviation (SD) between true and estimated ability parameter ( $\theta, \hat{\theta}$ ) and between true and estimated person slowness parameter ( $s, \hat{s}$ ) across 100 replications for 20 items and 1000 examinees

| | Using True<br>Item Parameters | Using Estimated<br>Item Parameters |
|---|---|---|
| Correlation of $\theta, \hat{\theta}$ (SD) | 0.8881 (0.0060) | 0.8873 (0.0060) |
| Correlation of $s, \hat{s}$ (SD) | 0.9384 (0.0036) | 0.9380 (0.0038) |

The distributions of the correlations for ability and person slowness parameters are illustrated in the histograms in Figure C.1 and C.2 respectively. For ability parameters, the bell-shaped histograms show many observations around the mean of the correlations given in Table 4.13. For person slowness parameters, the histograms given in Table 4.13 are slightly left skewed, which indicates that many correlation values are high. The high values of correlations and histograms indicate that the correlations are relatively consistent and precise.

Therefore, it is concluded that ability and person slowness parameters are reasonably well estimated using the MAP procedure, not only when the true item parameters are known, but also when item parameters are estimated.

# CHAPTER 5

# DISCUSSION

## Summary

The results derived in the preceding chapters suggest that the joint distribution of the item responses and response times is an effective approach. The estimation procedures work well and produce reasonably accurate parameter estimates. In particular, the findings based on the marginal maximum likelihood (MML) procedure employed to estimate the item parameters, as well as the maximum a posteriori (MAP) procedure implemented to estimate person parameters, were consistent and accurate. Overall, parameter recovery in the simulations was substantial. Hence, the extended IRT model incorporating response time suggested in this work is a successful model.

In the first simulation study, it was found that the MML method used in this study was successful at parameter estimation across all four simulation conditions, which were the combinations of 20 and 40 items with 1000 and 2000 examinees. The item parameters and additional parameters were estimated fairly well when the response accuracy and response time data were generated from the assumed model with discrete distributions of ability and person slowness parameters. The bias, SE and RMSE were reasonably small across all four simulation conditions. The observed variances of the estimates indicated that the parameter average estimates based on 100 replications were consistent and accurate across four simulation conditions. In general, the greater the number of examinees and the longer the test length, the better the parameters estimates. However, it was found that the impact of number of examinees on the parameter average estimates was larger than the impact of test length.

The findings of the second simulation were that the maximum a posteriori (MAP) procedure effectively estimated the continuous ability and person slowness parameters when the true item parameters were known. The means and standard deviations of two person parameters across all four simulation conditions were fairly close to the assumed true values. The small variability indicates that obtained means were consistent and pretty accurate. This was confirmed by the high and consistent values of correlations with correspondingly small variability. In general, the longer the test length, the better the correlations. It was found that the impact of test length on the average correlation estimates was larger than the impact of number of examinees. The standard deviations of correlations got smaller as the number of examinees and the number of items increased. However, it was discovered that the impact of number of examinees on the standard deviations of correlations was larger than the impact of test length.

As the third simulation study demonstrated, ability and person slowness parameter estimations were successfully estimated using the MAP procedure not only when the true item parameters were known but also when item parameters were estimated for a simulation condition of 20 items and 1000 examinees. Comparing the correlations of person parameters based on known true item parameters and on estimated item parameters, the correlations of person parameters using true item parameter case were almost same as those of the estimated item parameter case. Means and standard deviations of person parameters were comparable and their corresponding variability seemed to be very small. The high correlation values and stable distributions of the correlations indicated that the obtained correlations for person parameters were relatively consistent and precise.

## Limitations and Future Research

This work leaves a number of issues for future research. It should be noted that this study does not explicitly delve into the model fit issue due to its limited scope. Future research may focus on the test of the model fit. Specifically, a model fit statistic should be developed for this model to assess the model fit. From a practical perspective, the suggested model should be applied to real test data with response time at the item level, to evaluate how well the model fits the data.

Future research is also needed in validity studies in order to address the utility of the scores from response time models. The impact of such a utility of the scores has to be carefully investigated.

Some cautions should be mentioned with respect to the assumption of the discrete normal distribution for person parameters. In the present study, due to the straightforward model implementation, the ability and person slowness parameters are simulated based on discrete normal distributions. In other words, sampling the data from discrete and fitting the model to the discrete results in the finite sum, and the ease of integration for marginal maximum likelihood method which enables the ability parameter to be estimated easier than when the continuous distribution is assumed. In order to more closely represent the continuous normal distribution, it is recommended to increase the number of discrete spacing values to 40 or 80. However, this will create substantial computational burdens. In addition, a sequence of explorations could be also undertaken to investigate item parameter recovery assuming continuous ability parameters.

Simultaneously modeling response accuracy and response time is an intriguing but challenging endeavor. There are various ways to model the joint distribution of response accuracy and response time. Modeling the response time distribution in relation to other person and item parameters is a complicated undertaking. In particular, the relationship between ability and person slowness is unknown to us. In this study, the independence is assumed. Further work on alternative models is needed to consider this issue more thoroughly. Of particular of interest would be to investigate a different log linear model within a lognormal distribution family which assumes various correlations between ability and person slowness parameters.

## Conclusions

The joint distribution of item response and response time suggested in this work incorporates an important source of educational measurement data, examinee’s response time, which became available by computerized testing. This work attempted to improve on current IRT models that do not account for the response time when there is a time limit in the real testing context. As a consequence, on behalf of fairness and equity, the proposed model incorporating response time is an important step in the field to improve measurement quality.

In conclusion, the extended IRT model incorporating response time suggested in this work is a successful model. Overall, the model performance and the parameter recovery in the simulations were satisfactory. In summary, the suggested model and parameter estimation procedures are promising and represent a unique contribution to the field. Furthermore, this work should be viewed as a noteworthy step in exploring appropriate response time models and parameter estimation procedures. Obviously much work remains.

## Practical Implications

The examinee behavior in testing is usually observed by the examinee's test score based on the accuracy of the test items (Schnipke & Scrams, 2002). That is to say, in psychometrics, the relatively simple scores are derived from classical test theory or item response theory. This is a rather narrow perspective about how examinees process information. Traditionally, among cognitive psychologists, response time has been considered as a source of information about how the mind processes information. The developed model in this work is a scoring model which uses response time in the scoring process. Therefore, this model will offer benefits to the field of psychometrics and provide a more accurate estimation of the examinee's ability.

# APPENDIX A

## DETAILED ANALYSIS OF SIMULATION DESIGN 1 – ITEM PARAMETER ESTIMATION USING MML

The results of parameter recoveries for three additional simulation conditions are presented. They involve 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees.

Tables A.1, A.9, and A.17 include bias, SE and RMSE and average estimates based on 100 replications of the item discrimination parameters ($a$) for 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. Tables A.3, A.11, and A.19 show bias, SE and RMSE and average estimates based on 100 replications of item difficulty parameter ($b$) for 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. Tables A.5, A.13, and A.21 show bias, SE and RMSE and average estimates based on 100 replications of item slowness parameter ($r$) for 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. The additional model parameter ($\eta, g, \sigma, \sigma_s$) estimates are contained in Tables A.7, A.15, and A.23 for the conditions with 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively.

In addition, consistency is assessed based on sample variances of the parameter estimates ($S^2$) and means of the variance estimates from the Fisher information ($\overline{I}$) and corresponding standard deviations (SD) over 100 simulation replications. Tables A.2, A.10, and A.18 display the consistency checks of the item discrimination parameters ($a$) for the conditions with20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. Tables A.4, A.12, and A.20 display the consistency checks of the item difficulty parameter ($b$) for the conditions with 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. Tables A.6, A.14, and A.22 exhibit the consistency checks of the item slowness parameter ( $r$ ) for the conditions with 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively. The consistency check of the additional model parameter ( $\eta, g, \sigma, \sigma_s$ ) estimates are contained in Tables A.8, A.16, and A.24 for the conditions with 20 items and 2000 examinees, 40 items and 1000 examinees, and 40 items and 2000 examinees, respectively.

Table A.1 Across-item bias, SE and RMSE of the item discrimination parameter $(a)$ average estimates based on 100 replications for 20 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 1.0036 | 0.9989 | 0.0554 | -0.0047 | 0.0553 |
| 2 | 1.0650 | 1.0662 | 0.0565 | 0.0012 | 0.0563 |
| 3 | 0.9497 | 0.9416 | 0.0517 | -0.0081 | 0.0521 |
| 4 | 1.2234 | 1.2297 | 0.0485 | 0.0063 | 0.0486 |
| 5 | 0.9407 | 0.9384 | 0.0491 | -0.0023 | 0.0489 |
| 6 | 0.6576 | 0.6517 | 0.0435 | -0.0059 | 0.0436 |
| 7 | 0.8279 | 0.8203 | 0.0499 | -0.0076 | 0.0502 |
| 8 | 0.9844 | 0.9800 | 0.0500 | -0.0044 | 0.0499 |
| 9 | 1.5805 | 1.5780 | 0.0598 | -0.0025 | 0.0596 |
| 10 | 1.2155 | 1.2092 | 0.0523 | -0.0063 | 0.0525 |
| 11 | 0.7860 | 0.7817 | 0.0492 | -0.0043 | 0.0491 |
| 12 | 1.3716 | 1.3692 | 0.0497 | -0.0024 | 0.0495 |
| 13 | 1.2027 | 1.2116 | 0.0542 | 0.0089 | 0.0547 |
| 14 | 1.5569 | 1.5529 | 0.0569 | -0.0040 | 0.0568 |
| 15 | 1.2431 | 1.2470 | 0.0591 | 0.0039 | 0.0589 |
| 16 | 0.6453 | 0.6481 | 0.0460 | 0.0028 | 0.0459 |
| 17 | 1.4352 | 1.4415 | 0.0649 | 0.0063 | 0.0649 |
| 18 | 0.8156 | 0.8190 | 0.0450 | 0.0034 | 0.0449 |
| 19 | 1.7798 | 1.7709 | 0.0641 | -0.0089 | 0.0644 |
| 20 | 0.9110 | 0.9153 | 0.0548 | 0.0043 | 0.0547 |

Table A.2 Sample variance of the item discrimination ($a$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_a$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_a$ (SD) |
|---|---|---|
| 1 | 0.003070 | 0.002889 (0.000172) |
| 2 | 0.003197 | 0.002909 (0.000171) |
| 3 | 0.002676 | 0.002614 (0.000145) |
| 4 | 0.002349 | 0.003045 (0.000173) |
| 5 | 0.002413 | 0.002466 (0.000131) |
| 6 | 0.001889 | 0.002119 (0.000093) |
| 7 | 0.002486 | 0.002235 (0.000108) |
| 8 | 0.002499 | 0.002377 (0.000125) |
| 9 | 0.003576 | 0.003552 (0.000237) |
| 10 | 0.002740 | 0.002679 (0.000150) |
| 11 | 0.002421 | 0.002132 (0.000097) |
| 12 | 0.002471 | 0.002997 (0.000165) |
| 13 | 0.002938 | 0.002716 (0.000156) |
| 14 | 0.003241 | 0.003549 (0.000229) |
| 15 | 0.003493 | 0.002895 (0.000178) |
| 16 | 0.002116 | 0.002144 (0.000098) |
| 17 | 0.004209 | 0.003523 (0.000269) |
| 18 | 0.002027 | 0.002425 (0.000120) |
| 19 | 0.004103 | 0.004868 (0.000349) |
| 20 | 0.003003 | 0.002713 (0.000148) |

Table A.3 Across-item bias, SE and RMSE of the item difficulty parameter ($b$) average estimates based on 100 replications for 20 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | -2.0000 | -2.0128 | 0.1099 | -0.0128 | 0.1101 |
| 2 | -1.7895 | -1.7956 | 0.0975 | -0.0061 | 0.0972 |
| 3 | -1.5789 | -1.5881 | 0.0926 | -0.0092 | 0.0926 |
| 4 | -1.3684 | -1.3658 | 0.0638 | 0.0026 | 0.0636 |
| 5 | -1.1579 | -1.1705 | 0.0736 | -0.0127 | 0.0743 |
| 6 | -0.9474 | -0.9401 | 0.0925 | 0.0073 | 0.0923 |
| 7 | -0.7368 | -0.7513 | 0.0740 | -0.0144 | 0.0750 |
| 8 | -0.5263 | -0.5342 | 0.0602 | -0.0079 | 0.0605 |
| 9 | -0.3158 | -0.3101 | 0.0410 | 0.0057 | 0.0412 |
| 10 | -0.1053 | -0.1028 | 0.0439 | 0.0024 | 0.0437 |
| 11 | 0.1053 | 0.1120 | 0.0606 | 0.0067 | 0.0607 |
| 12 | 0.3158 | 0.3183 | 0.0409 | 0.0025 | 0.0408 |
| 13 | 0.5263 | 0.5300 | 0.0525 | 0.0036 | 0.0524 |
| 14 | 0.7368 | 0.7427 | 0.0455 | 0.0058 | 0.0456 |
| 15 | 0.9474 | 0.9508 | 0.0542 | 0.0034 | 0.0540 |
| 16 | 1.1579 | 1.1570 | 0.1157 | -0.0009 | 0.1152 |
| 17 | 1.3684 | 1.3724 | 0.0607 | 0.0040 | 0.0605 |
| 18 | 1.5789 | 1.5802 | 0.1109 | 0.0012 | 0.1103 |
| 19 | 1.7895 | 1.7971 | 0.0634 | 0.0076 | 0.0635 |
| 20 | 2.0000 | 1.9878 | 0.1185 | -0.0122 | 0.1185 |

Table A.4 Sample variance of the item difficulty ($b$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_b$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_b$ (SD) |
|---|---|---|
| 1 | 0.012071 | 0.011872 (0.002317) |
| 2 | 0.009504 | 0.008680 (0.001477) |
| 3 | 0.008578 | 0.008725 (0.001501) |
| 4 | 0.004076 | 0.004584 (0.000457) |
| 5 | 0.005412 | 0.005888 (0.000840) |
| 6 | 0.008556 | 0.009269 (0.001685) |
| 7 | 0.005469 | 0.005202 (0.000769) |
| 8 | 0.003629 | 0.003336 (0.000320) |
| 9 | 0.001681 | 0.001615 (0.000073) |
| 10 | 0.001927 | 0.002129 (0.000128) |
| 11 | 0.003676 | 0.004096 (0.000436) |
| 12 | 0.001676 | 0.001891 (0.000091) |
| 13 | 0.002757 | 0.002449 (0.000193) |
| 14 | 0.002071 | 0.002012 (0.000124) |
| 15 | 0.002934 | 0.003148 (0.000313) |
| 16 | 0.013395 | 0.011765 (0.002877) |
| 17 | 0.003683 | 0.003730 (0.000383) |
| 18 | 0.012292 | 0.011272 (0.002327) |
| 19 | 0.004014 | 0.004510 (0.000450) |
| 20 | 0.014045 | 0.013638 (0.002906) |

Table A.5 Across-item bias, SE and RMSE of the item slowness parameter ( $r$ ) average estimates based on 100 replications for 20 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 0.7055 | 0.7081 | 0.0597 | 0.0026 | 0.0594 |
| 2 | 0.5795 | 0.5801 | 0.0606 | 0.0006 | 0.0603 |
| 3 | 0.5019 | 0.5064 | 0.0527 | 0.0045 | 0.0526 |
| 4 | 1.0000 | 1.0003 | 0.0544 | 0.0003 | 0.0542 |
| 5 | 0.8145 | 0.8132 | 0.0445 | -0.0013 | 0.0443 |
| 6 | 0.5445 | 0.5539 | 0.0383 | 0.0094 | 0.0393 |
| 7 | 0.8626 | 0.8655 | 0.0390 | 0.0029 | 0.0389 |
| 8 | 0.7350 | 0.7340 | 0.0445 | -0.0010 | 0.0443 |
| 9 | 0.8714 | 0.8793 | 0.0440 | 0.0079 | 0.0445 |
| 10 | 0.9496 | 0.9556 | 0.0389 | 0.0060 | 0.0392 |
| 11 | 0.5249 | 0.5302 | 0.0337 | 0.0053 | 0.0339 |
| 12 | 0.5350 | 0.5330 | 0.0351 | -0.0020 | 0.0350 |
| 13 | 0.6874 | 0.6912 | 0.0372 | 0.0038 | 0.0372 |
| 14 | 0.6227 | 0.6280 | 0.0442 | 0.0053 | 0.0443 |
| 15 | 0.8326 | 0.8378 | 0.0502 | 0.0052 | 0.0502 |
| 16 | 0.8298 | 0.8333 | 0.0456 | 0.0035 | 0.0455 |
| 17 | 0.7892 | 0.7991 | 0.0539 | 0.0099 | 0.0545 |
| 18 | 0.9110 | 0.9129 | 0.0501 | 0.0019 | 0.0499 |
| 19 | 0.6951 | 0.6930 | 0.0815 | -0.0021 | 0.0812 |
| 20 | 0.5439 | 0.5398 | 0.0525 | -0.0041 | 0.0524 |

Table A.6 Sample variance of the item slowness ($r$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_r$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_r$ (SD) |
|---|---|---|
| 1 | 0.003560 | 0.003729 (0.000271) |
| 2 | 0.003676 | 0.003464 (0.000245) |
| 3 | 0.002778 | 0.002667 (0.000174) |
| 4 | 0.002964 | 0.002961 (0.000201) |
| 5 | 0.001978 | 0.002084 (0.000115) |
| 6 | 0.001469 | 0.001577 (0.000071) |
| 7 | 0.001521 | 0.001602 (0.000076) |
| 8 | 0.001979 | 0.001560 (0.000071) |
| 9 | 0.001934 | 0.001648 (0.000069) |
| 10 | 0.001515 | 0.001460 (0.000055) |
| 11 | 0.001136 | 0.001375 (0.000049) |
| 12 | 0.001234 | 0.001561 (0.000064) |
| 13 | 0.001387 | 0.001647 (0.000074) |
| 14 | 0.001956 | 0.002092 (0.000111) |
| 15 | 0.002519 | 0.002149 (0.000114) |
| 16 | 0.002075 | 0.001675 (0.000082) |
| 17 | 0.002903 | 0.003334 (0.000193) |
| 18 | 0.002512 | 0.002330 (0.000142) |
| 19 | 0.006648 | 0.005921 (0.000422) |
| 20 | 0.002761 | 0.003277 (0.000213) |

Table A.7 Across-item bias, SE and RMSE of the parameter $(\eta, g, \sigma, \sigma_s)$ average estimates based on 100 replications for 20 items and 2000 examinees

| Parameter | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| $\eta$ | 0.002 | 0.0019 | 0.0007 | -0.00004 | 0.0007 |
| $g$ | 0.5 | 0.4999 | 0.0208 | -0.00002 | 0.0207 |
| $\sigma$ | 1 | 0.9998 | 0.0035 | -0.0001 | 0.0035 |
| $\sigma_s$ | 1 | 0.9983 | 0.0151 | -0.0016 | 0.0151 |

Table A.8 Sample variance of the parameter ($\eta, g, \sigma, \sigma_s$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\bar{I}$) and corresponding standard deviation (SD) over 100 simulation replications for 20 items and 2000 examinees

| Parameter | $S^2$ | $\bar{I}$ (SD) |
|---|---|---|
| $\eta$ | 0.00000051 | 0.00000056 (0.00000011) |
| $g$ | 0.00043302 | 0.00042891 (0.00003321) |
| $\sigma$ | 0.00001249 | 0.00001327 (0.00000009) |
| $\sigma_s$ | 0.00023050 | 0.00021384 (0.00000816) |

Table A.9 Across-item bias, SE and RMSE of the item discrimination parameter ($a$) average estimates based on 100 replications for 40 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 1.0036 | 1.0100 | 0.0671 | 0.0064 | 0.0670 |
| 2 | 1.0320 | 1.0378 | 0.0712 | 0.0058 | 0.0711 |
| 3 | 1.0650 | 1.0614 | 0.0686 | -0.0036 | 0.0684 |
| 4 | 0.9860 | 0.9890 | 0.0652 | 0.0030 | 0.0649 |
| 5 | 0.9497 | 0.9577 | 0.0655 | 0.0080 | 0.0656 |
| 6 | 1.1230 | 1.1164 | 0.0703 | -0.0066 | 0.0702 |
| 7 | 1.2234 | 1.2321 | 0.0623 | 0.0087 | 0.0626 |
| 8 | 0.9670 | 0.9672 | 0.0648 | 0.0002 | 0.0645 |
| 9 | 0.9407 | 0.9413 | 0.0608 | 0.0006 | 0.0605 |
| 10 | 0.8760 | 0.8753 | 0.0642 | -0.0007 | 0.0639 |
| 11 | 0.6576 | 0.6609 | 0.0587 | 0.0033 | 0.0585 |
| 12 | 0.7560 | 0.7533 | 0.0603 | -0.0027 | 0.0600 |
| 13 | 0.8279 | 0.8370 | 0.0583 | 0.0091 | 0.0587 |
| 14 | 0.9340 | 0.9385 | 0.0647 | 0.0045 | 0.0645 |
| 15 | 0.9844 | 0.9876 | 0.0746 | 0.0032 | 0.0743 |
| 16 | 1.2450 | 1.2434 | 0.0699 | -0.0016 | 0.0695 |
| 17 | 1.5805 | 1.5763 | 0.0798 | -0.0042 | 0.0795 |
| 18 | 1.4350 | 1.4341 | 0.0703 | -0.0009 | 0.0700 |
| 19 | 1.2155 | 1.2144 | 0.0746 | -0.0011 | 0.0743 |
| 20 | 0.9870 | 0.9837 | 0.0587 | -0.0033 | 0.0585 |
| 21 | 0.7860 | 0.7896 | 0.0524 | 0.0036 | 0.0523 |
| 22 | 1.1240 | 1.1334 | 0.0646 | 0.0094 | 0.0649 |
| 23 | 1.3716 | 1.3621 | 0.0676 | -0.0095 | 0.0679 |
| 24 | 1.1000 | 1.1032 | 0.0627 | 0.0032 | 0.0625 |
| 25 | 1.2027 | 1.2087 | 0.0619 | 0.0060 | 0.0619 |
| 26 | 1.3450 | 1.3424 | 0.0643 | -0.0026 | 0.0640 |
| 27 | 1.5569 | 1.5595 | 0.0753 | 0.0026 | 0.0750 |
| 28 | 1.4320 | 1.4263 | 0.0629 | -0.0057 | 0.0628 |
| 29 | 1.2431 | 1.2457 | 0.0745 | 0.0026 | 0.0742 |
| 30 | 0.8780 | 0.8785 | 0.0575 | 0.0005 | 0.0572 |
| 31 | 0.6453 | 0.6473 | 0.0536 | 0.0020 | 0.0533 |
| 32 | 1.2220 | 1.2303 | 0.0686 | 0.0083 | 0.0688 |
| 33 | 1.4352 | 1.4440 | 0.0753 | 0.0088 | 0.0754 |
| 34 | 0.7870 | 0.7828 | 0.0708 | -0.0042 | 0.0706 |
| 35 | 0.8156 | 0.8259 | 0.0601 | 0.0103 | 0.0606 |
| 36 | 1.5670 | 1.5636 | 0.0716 | -0.0034 | 0.0713 |
| 37 | 1.7798 | 1.7772 | 0.0903 | -0.0026 | 0.0899 |
| 38 | 1.4560 | 1.4479 | 0.0683 | -0.0081 | 0.0684 |
| 39 | 0.9110 | 0.9149 | 0.0727 | 0.0039 | 0.0725 |
| 40 | 0.7980 | 0.8036 | 0.0668 | 0.0056 | 0.0667 |

Table A.10 Sample variance of the item discrimination ($a$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_a$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_a$ (SD) |
|---|---|---|
| 1 | 0.004498 | 0.004906 (0.000310) |
| 2 | 0.005073 | 0.004928 (0.000321) |
| 3 | 0.004707 | 0.004899 (0.000313) |
| 4 | 0.004249 | 0.004633 (0.000325) |
| 5 | 0.004284 | 0.004503 (0.000281) |
| 6 | 0.004936 | 0.004857 (0.000331) |
| 7 | 0.003884 | 0.005070 (0.000299) |
| 8 | 0.004198 | 0.004350 (0.000264) |
| 9 | 0.003702 | 0.004233 (0.000261) |
| 10 | 0.004126 | 0.004058 (0.000252) |
| 11 | 0.003445 | 0.003683 (0.000187) |
| 12 | 0.003633 | 0.003791 (0.000196) |
| 13 | 0.003399 | 0.003886 (0.000212) |
| 14 | 0.004188 | 0.004024 (0.000236) |
| 15 | 0.005569 | 0.004091 (0.000263) |
| 16 | 0.004880 | 0.004642 (0.000303) |
| 17 | 0.006371 | 0.005691 (0.000435) |
| 18 | 0.004947 | 0.005158 (0.000369) |
| 19 | 0.005569 | 0.004518 (0.000322) |
| 20 | 0.003441 | 0.004000 (0.000225) |
| 21 | 0.002745 | 0.003699 (0.000180) |
| 22 | 0.004172 | 0.004309 (0.000287) |
| 23 | 0.004569 | 0.004916 (0.000348) |
| 24 | 0.003935 | 0.004265 (0.000246) |
| 25 | 0.003830 | 0.004545 (0.000307) |
| 26 | 0.004134 | 0.004947 (0.000326) |
| 27 | 0.005670 | 0.005712 (0.000439) |
| 28 | 0.003956 | 0.005289 (0.000346) |
| 29 | 0.005557 | 0.004795 (0.000336) |
| 30 | 0.003310 | 0.004009 (0.000245) |
| 31 | 0.002869 | 0.003694 (0.000184) |
| 32 | 0.004707 | 0.004914 (0.000311) |
| 33 | 0.005670 | 0.005673 (0.000400) |
| 34 | 0.005014 | 0.004018 (0.000217) |
| 35 | 0.003606 | 0.004142 (0.000246) |
| 36 | 0.005121 | 0.006404 (0.000513) |
| 37 | 0.008149 | 0.007493 (0.000649) |
| 38 | 0.004666 | 0.006128 (0.000471) |
| 39 | 0.005287 | 0.004583 (0.000306) |
| 40 | 0.004460 | 0.004397 (0.000265) |

Table A.11 Across-item bias, SE and RMSE of the item difficulty parameter ( $b$ ) average estimates based on 100 replications for 40 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | -2.0000 | -1.9935 | 0.1479 | 0.0065 | 0.1473 |
| 2 | -1.8974 | -1.9099 | 0.1346 | -0.0124 | 0.1345 |
| 3 | -1.7949 | -1.8049 | 0.1299 | -0.0101 | 0.1297 |
| 4 | -1.6923 | -1.6830 | 0.1142 | 0.0093 | 0.1140 |
| 5 | -1.5897 | -1.6039 | 0.1291 | -0.0142 | 0.1292 |
| 6 | -1.4872 | -1.5220 | 0.0999 | -0.0348 | 0.1053 |
| 7 | -1.3846 | -1.3675 | 0.0818 | 0.0171 | 0.0831 |
| 8 | -1.2821 | -1.2901 | 0.1213 | -0.0081 | 0.1210 |
| 9 | -1.1795 | -1.1880 | 0.1142 | -0.0085 | 0.1140 |
| 10 | -1.0769 | -1.0831 | 0.1090 | -0.0061 | 0.1087 |
| 11 | -0.9744 | -0.9779 | 0.1253 | -0.0035 | 0.1247 |
| 12 | -0.8718 | -0.8895 | 0.1319 | -0.0177 | 0.1324 |
| 13 | -0.7692 | -0.7846 | 0.1012 | -0.0154 | 0.1019 |
| 14 | -0.6667 | -0.6746 | 0.0820 | -0.0079 | 0.0820 |
| 15 | -0.5641 | -0.5696 | 0.0842 | -0.0055 | 0.0839 |
| 16 | -0.4615 | -0.4584 | 0.0673 | 0.0031 | 0.0670 |
| 17 | -0.3590 | -0.3627 | 0.0586 | -0.0037 | 0.0584 |
| 18 | -0.2564 | -0.2618 | 0.0538 | -0.0054 | 0.0538 |
| 19 | -0.1538 | -0.1552 | 0.0554 | -0.0013 | 0.0551 |
| 20 | -0.0513 | -0.0435 | 0.0743 | 0.0078 | 0.0743 |
| 21 | 0.0513 | 0.0446 | 0.0857 | -0.0067 | 0.0856 |
| 22 | 0.1538 | 0.1492 | 0.0639 | -0.0047 | 0.0638 |
| 23 | 0.2564 | 0.2548 | 0.0582 | -0.0016 | 0.0579 |
| 24 | 0.3590 | 0.3511 | 0.0784 | -0.0078 | 0.0784 |
| 25 | 0.4615 | 0.4548 | 0.0629 | -0.0067 | 0.0630 |
| 26 | 0.5641 | 0.5670 | 0.0674 | 0.0029 | 0.0671 |
| 27 | 0.6667 | 0.6606 | 0.0562 | -0.0061 | 0.0563 |
| 28 | 0.7692 | 0.7706 | 0.0563 | 0.0013 | 0.0560 |
| 29 | 0.8718 | 0.8672 | 0.0808 | -0.0046 | 0.0806 |
| 30 | 0.9744 | 0.9768 | 0.0991 | 0.0024 | 0.0986 |
| 31 | 1.0769 | 1.0719 | 0.1427 | -0.0050 | 0.1421 |
| 32 | 1.1795 | 1.1728 | 0.0881 | -0.0067 | 0.0879 |
| 33 | 1.2821 | 1.2733 | 0.0861 | -0.0087 | 0.0861 |
| 34 | 1.3846 | 1.4089 | 0.1404 | 0.0243 | 0.1417 |
| 35 | 1.4872 | 1.4814 | 0.1343 | -0.0058 | 0.1338 |
| 36 | 1.5897 | 1.5877 | 0.0815 | -0.0020 | 0.0811 |
| 37 | 1.6923 | 1.6895 | 0.0871 | -0.0028 | 0.0867 |
| 38 | 1.7949 | 1.8000 | 0.0902 | 0.0052 | 0.0899 |
| 39 | 1.8974 | 1.8963 | 0.1436 | -0.0011 | 0.1429 |
| 40 | 2.0000 | 2.0099 | 0.1709 | 0.0099 | 0.1703 |

Table A.12 Sample variance of the item difficulty ($b$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_b$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_b$ (SD) |
|---|---|---|
| 1 | 0.021876 | 0.021057 (0.004994) |
| 2 | 0.018112 | 0.018683 (0.004293) |
| 3 | 0.016879 | 0.016284 (0.003546) |
| 4 | 0.013034 | 0.016266 (0.003198) |
| 5 | 0.016660 | 0.016056 (0.003674) |
| 6 | 0.009979 | 0.011481 (0.001988) |
| 7 | 0.006684 | 0.008549 (0.001127) |
| 8 | 0.014722 | 0.011826 (0.002422) |
| 9 | 0.013048 | 0.011293 (0.002138) |
| 10 | 0.011889 | 0.011733 (0.002314) |
| 11 | 0.015707 | 0.018059 (0.004800) |
| 12 | 0.017394 | 0.013187 (0.003140) |
| 13 | 0.010246 | 0.009984 (0.001788) |
| 14 | 0.006722 | 0.007615 (0.001084) |
| 15 | 0.007085 | 0.006614 (0.000963) |
| 16 | 0.004525 | 0.004436 (0.000417) |
| 17 | 0.003437 | 0.003181 (0.000212) |
| 18 | 0.002893 | 0.003450 (0.000213) |
| 19 | 0.003067 | 0.004190 (0.000365) |
| 20 | 0.005519 | 0.005595 (0.000508) |
| 21 | 0.007351 | 0.007951 (0.000872) |
| 22 | 0.004085 | 0.004593 (0.000380) |
| 23 | 0.003387 | 0.003658 (0.000238) |
| 24 | 0.006139 | 0.005043 (0.000507) |
| 25 | 0.003958 | 0.004599 (0.000364) |
| 26 | 0.004539 | 0.004206 (0.000345) |
| 27 | 0.003162 | 0.003652 (0.000280) |
| 28 | 0.003170 | 0.004362 (0.000338) |
| 29 | 0.006533 | 0.005651 (0.000731) |
| 30 | 0.009814 | 0.010751 (0.001753) |
| 31 | 0.020368 | 0.020803 (0.006353) |
| 32 | 0.007765 | 0.007308 (0.001035) |
| 33 | 0.007406 | 0.006353 (0.000842) |
| 34 | 0.019699 | 0.019734 (0.005730) |
| 35 | 0.018038 | 0.018858 (0.004456) |
| 36 | 0.006646 | 0.007662 (0.000887) |
| 37 | 0.007579 | 0.007358 (0.000960) |
| 38 | 0.008141 | 0.010433 (0.001493) |
| 39 | 0.020630 | 0.023078 (0.006101) |
| 40 | 0.029195 | 0.032237 (0.009373) |

Table A.13 Across-item bias, SE and RMSE of the item slowness parameter ( $r$ ) average estimates based on 100 replications for 40 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 0.7055 | 0.6956 | 0.0832 | -0.0099 | 0.0833 |
| 2 | 0.6760 | 0.6552 | 0.0788 | -0.0208 | 0.0811 |
| 3 | 0.5795 | 0.5787 | 0.0802 | -0.0008 | 0.0798 |
| 4 | 0.5340 | 0.5322 | 0.0591 | -0.0018 | 0.0589 |
| 5 | 0.5019 | 0.4787 | 0.0691 | -0.0232 | 0.0726 |
| 6 | 0.8980 | 0.8825 | 0.0645 | -0.0155 | 0.0660 |
| 7 | 1.0000 | 0.9989 | 0.0803 | -0.0011 | 0.0799 |
| 8 | 0.9230 | 0.9151 | 0.0662 | -0.0079 | 0.0663 |
| 9 | 0.8145 | 0.8010 | 0.0652 | -0.0135 | 0.0662 |
| 10 | 0.7860 | 0.7782 | 0.0573 | -0.0078 | 0.0575 |
| 11 | 0.5445 | 0.5340 | 0.0517 | -0.0105 | 0.0525 |
| 12 | 0.6780 | 0.6727 | 0.0547 | -0.0053 | 0.0547 |
| 13 | 0.8626 | 0.8479 | 0.0601 | -0.0147 | 0.0616 |
| 14 | 0.7920 | 0.7806 | 0.0593 | -0.0114 | 0.0601 |
| 15 | 0.7350 | 0.7286 | 0.0540 | -0.0064 | 0.0541 |
| 16 | 0.8120 | 0.8161 | 0.0556 | 0.0041 | 0.0555 |
| 17 | 0.8714 | 0.8712 | 0.0609 | -0.0002 | 0.0606 |
| 18 | 0.9230 | 0.9162 | 0.0633 | -0.0068 | 0.0633 |
| 19 | 0.9496 | 0.9467 | 0.0556 | -0.0029 | 0.0554 |
| 20 | 0.6780 | 0.6766 | 0.0458 | -0.0014 | 0.0456 |
| 21 | 0.5249 | 0.5196 | 0.0556 | -0.0053 | 0.0556 |
| 22 | 0.5950 | 0.5889 | 0.0554 | -0.0061 | 0.0555 |
| 23 | 0.5350 | 0.5349 | 0.0532 | -0.0001 | 0.0530 |
| 24 | 0.6230 | 0.6191 | 0.0581 | -0.0039 | 0.0579 |
| 25 | 0.6874 | 0.6836 | 0.0536 | -0.0038 | 0.0535 |
| 26 | 0.6920 | 0.6935 | 0.0562 | 0.0015 | 0.0559 |
| 27 | 0.6227 | 0.6190 | 0.0591 | -0.0037 | 0.0589 |
| 28 | 0.7970 | 0.7980 | 0.0618 | 0.0010 | 0.0615 |
| 29 | 0.8326 | 0.8338 | 0.0633 | 0.0012 | 0.0630 |
| 30 | 0.8010 | 0.8039 | 0.0596 | 0.0029 | 0.0594 |
| 31 | 0.8298 | 0.8316 | 0.0567 | 0.0018 | 0.0564 |
| 32 | 0.7230 | 0.7282 | 0.0641 | 0.0052 | 0.0640 |
| 33 | 0.7892 | 0.7916 | 0.0694 | 0.0024 | 0.0691 |
| 34 | 0.8930 | 0.8982 | 0.0647 | 0.0052 | 0.0646 |
| 35 | 0.9110 | 0.9179 | 0.0610 | 0.0069 | 0.0611 |
| 36 | 0.7520 | 0.7618 | 0.0914 | 0.0098 | 0.0914 |
| 37 | 0.6951 | 0.6993 | 0.0946 | 0.0042 | 0.0942 |
| 38 | 0.5930 | 0.5929 | 0.0786 | -0.0001 | 0.0782 |
| 39 | 0.5439 | 0.5466 | 0.0662 | 0.0027 | 0.0660 |
| 40 | 0.6250 | 0.6367 | 0.0731 | 0.0117 | 0.0737 |

Table A.14 Sample variance of the item slowness ($r$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_r$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 1000 examinees

| Item | $S^2$ | $\overline{I}_r$ (SD) |
|---|---|---|
| 1 | 0.006918 | 0.006142 (0.000586) |
| 2 | 0.006208 | 0.005995 (0.000518) |
| 3 | 0.006431 | 0.005761 (0.000497) |
| 4 | 0.003498 | 0.005007 (0.000345) |
| 5 | 0.004781 | 0.004662 (0.000360) |
| 6 | 0.004163 | 0.005089 (0.000405) |
| 7 | 0.006441 | 0.005030 (0.000417) |
| 8 | 0.004381 | 0.004006 (0.000297) |
| 9 | 0.004248 | 0.003744 (0.000262) |
| 10 | 0.003281 | 0.003440 (0.000218) |
| 11 | 0.002676 | 0.002991 (0.000168) |
| 12 | 0.002991 | 0.003039 (0.000172) |
| 13 | 0.003614 | 0.003046 (0.000162) |
| 14 | 0.003514 | 0.003047 (0.000157) |
| 15 | 0.002917 | 0.002989 (0.000141) |
| 16 | 0.003092 | 0.003077 (0.000152) |
| 17 | 0.003710 | 0.003203 (0.000165) |
| 18 | 0.004006 | 0.003020 (0.000141) |
| 19 | 0.003089 | 0.002851 (0.000116) |
| 20 | 0.002095 | 0.002736 (0.000116) |
| 21 | 0.003096 | 0.002674 (0.000113) |
| 22 | 0.003069 | 0.002805 (0.000125) |
| 23 | 0.002832 | 0.002954 (0.000140) |
| 24 | 0.003375 | 0.002883 (0.000139) |
| 25 | 0.002874 | 0.003030 (0.000153) |
| 26 | 0.003155 | 0.003266 (0.000185) |
| 27 | 0.003495 | 0.003636 (0.000219) |
| 28 | 0.003821 | 0.003724 (0.000240) |
| 29 | 0.004011 | 0.003674 (0.000250) |
| 30 | 0.003551 | 0.003287 (0.000213) |
| 31 | 0.003210 | 0.003026 (0.000171) |
| 32 | 0.004114 | 0.004369 (0.000364) |
| 33 | 0.004811 | 0.005268 (0.000439) |
| 34 | 0.004187 | 0.003663 (0.000260) |
| 35 | 0.003725 | 0.003905 (0.000283) |
| 36 | 0.008346 | 0.007221 (0.000655) |
| 37 | 0.008945 | 0.008923 (0.000819) |
| 38 | 0.006180 | 0.007916 (0.000625) |
| 39 | 0.004388 | 0.005197 (0.000396) |
| 40 | 0.005342 | 0.004869 (0.000429) |

Table A.15 Across-item bias, SE and RMSE of the parameter $(\eta, g, \sigma, \sigma_s)$ average estimates based on 100 replications for 40 items and 1000 examinees

| Parameter | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| $\eta$ | 0.002 | 0.0019 | 0.0008 | -0.0001 | 0.0008 |
| $g$ | 0.5 | 0.5045 | 0.0230 | 0.0045 | 0.0233 |
| $\sigma$ | 1 | 0.9997 | 0.0036 | -0.0003 | 0.0036 |
| $\sigma_s$ | 1 | 0.9978 | 0.0169 | -0.0022 | 0.0170 |

Table A.16 Sample variance of the parameter ($\eta, g, \sigma, \sigma_s$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\bar{I}$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 1000 examinees

| Parameter | $S^2$ | $\bar{I}$ (SD) |
|---|---|---|
| $\eta$ | 0.00000059 | 0.00000056 (0.00000012) |
| $g$ | 0.00052898 | 0.00053307 (0.00005824) |
| $\sigma$ | 0.00001299 | 0.00001287 (0.00000009) |
| $\sigma_s$ | 0.00028720 | 0.00028950 (0.00001715) |

Table A.17 Across-item bias, SE and RMSE of the item discrimination parameter ($a$) average estimates based on 100 replications for 40 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 1.0036 | 1.0073 | 0.0468 | 0.0037 | 0.0467 |
| 2 | 1.0320 | 1.0267 | 0.0488 | -0.0053 | 0.0488 |
| 3 | 1.0650 | 1.0742 | 0.0488 | 0.0092 | 0.0494 |
| 4 | 0.9860 | 0.9927 | 0.0448 | 0.0067 | 0.0451 |
| 5 | 0.9497 | 0.9475 | 0.0508 | -0.0022 | 0.0505 |
| 6 | 1.1230 | 1.1194 | 0.0498 | -0.0036 | 0.0497 |
| 7 | 1.2234 | 1.2279 | 0.0537 | 0.0045 | 0.0536 |
| 8 | 0.9670 | 0.9693 | 0.0514 | 0.0023 | 0.0512 |
| 9 | 0.9407 | 0.9440 | 0.0481 | 0.0033 | 0.0479 |
| 10 | 0.8760 | 0.8763 | 0.0502 | 0.0003 | 0.0500 |
| 11 | 0.6576 | 0.6562 | 0.0378 | -0.0014 | 0.0376 |
| 12 | 0.7560 | 0.7616 | 0.0442 | 0.0056 | 0.0443 |
| 13 | 0.8279 | 0.8285 | 0.0451 | 0.0006 | 0.0448 |
| 14 | 0.9340 | 0.9352 | 0.0411 | 0.0012 | 0.0410 |
| 15 | 0.9844 | 0.9859 | 0.0450 | 0.0015 | 0.0448 |
| 16 | 1.2450 | 1.2505 | 0.0425 | 0.0055 | 0.0427 |
| 17 | 1.5805 | 1.5869 | 0.0460 | 0.0064 | 0.0462 |
| 18 | 1.4350 | 1.4441 | 0.0463 | 0.0091 | 0.0469 |
| 19 | 1.2155 | 1.2253 | 0.0552 | 0.0098 | 0.0558 |
| 20 | 0.9870 | 0.9895 | 0.0442 | 0.0025 | 0.0441 |
| 21 | 0.7860 | 0.7881 | 0.0422 | 0.0021 | 0.0420 |
| 22 | 1.1240 | 1.1221 | 0.0419 | -0.0019 | 0.0418 |
| 23 | 1.3716 | 1.3737 | 0.0463 | 0.0021 | 0.0461 |
| 24 | 1.1000 | 1.1046 | 0.0529 | 0.0046 | 0.0528 |
| 25 | 1.2027 | 1.1984 | 0.0465 | -0.0043 | 0.0464 |
| 26 | 1.3450 | 1.3438 | 0.0486 | -0.0012 | 0.0484 |
| 27 | 1.5569 | 1.5611 | 0.0521 | 0.0042 | 0.0520 |
| 28 | 1.4320 | 1.4420 | 0.0554 | 0.0100 | 0.0560 |
| 29 | 1.2431 | 1.2473 | 0.0483 | 0.0042 | 0.0482 |
| 30 | 0.8780 | 0.8784 | 0.0435 | 0.0004 | 0.0433 |
| 31 | 0.6453 | 0.6427 | 0.0480 | -0.0026 | 0.0478 |
| 32 | 1.2220 | 1.2271 | 0.0506 | 0.0051 | 0.0506 |
| 33 | 1.4352 | 1.4472 | 0.0476 | 0.0120 | 0.0489 |
| 34 | 0.7870 | 0.7894 | 0.0438 | 0.0024 | 0.0436 |
| 35 | 0.8156 | 0.8163 | 0.0514 | 0.0007 | 0.0511 |
| 36 | 1.5670 | 1.5697 | 0.0598 | 0.0027 | 0.0596 |
| 37 | 1.7798 | 1.7786 | 0.0663 | -0.0012 | 0.0660 |
| 38 | 1.4560 | 1.4601 | 0.0522 | 0.0041 | 0.0521 |
| 39 | 0.9110 | 0.9169 | 0.0468 | 0.0059 | 0.0469 |
| 40 | 0.7980 | 0.7935 | 0.0395 | -0.0045 | 0.0396 |

Table A.18 Sample variance of the item discrimination ( $a$ ) estimates ( $S^2$ ) and mean of the variance estimates based on the Fisher information ( $\overline{I}_a$ ) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_a$ (SD) |
|---|---|---|
| 1 | 0.002189 | 0.002489 (0.000104) |
| 2 | 0.002381 | 0.002486 (0.000115) |
| 3 | 0.002383 | 0.002502 (0.000109) |
| 4 | 0.002005 | 0.002359 (0.000100) |
| 5 | 0.002576 | 0.002271 (0.000102) |
| 6 | 0.002479 | 0.002447 (0.000110) |
| 7 | 0.002880 | 0.002567 (0.000126) |
| 8 | 0.002639 | 0.002203 (0.000084) |
| 9 | 0.002309 | 0.002151 (0.000093) |
| 10 | 0.002521 | 0.002053 (0.000071) |
| 11 | 0.001427 | 0.001864 (0.000061) |
| 12 | 0.001951 | 0.001922 (0.000074) |
| 13 | 0.002030 | 0.001957 (0.000069) |
| 14 | 0.001693 | 0.002035 (0.000075) |
| 15 | 0.002026 | 0.002067 (0.000078) |
| 16 | 0.001810 | 0.002355 (0.000089) |
| 17 | 0.002116 | 0.002889 (0.000138) |
| 18 | 0.002140 | 0.002627 (0.000123) |
| 19 | 0.003050 | 0.002300 (0.000110) |
| 20 | 0.001956 | 0.002032 (0.000078) |
| 21 | 0.001779 | 0.001874 (0.000068) |
| 22 | 0.001759 | 0.002169 (0.000083) |
| 23 | 0.002140 | 0.002508 (0.000115) |
| 24 | 0.002795 | 0.002166 (0.000101) |
| 25 | 0.002159 | 0.002290 (0.000101) |
| 26 | 0.002364 | 0.002506 (0.000112) |
| 27 | 0.002718 | 0.002895 (0.000137) |
| 28 | 0.003067 | 0.002705 (0.000138) |
| 29 | 0.002332 | 0.002435 (0.000117) |
| 30 | 0.001894 | 0.002032 (0.000073) |
| 31 | 0.002301 | 0.001872 (0.000065) |
| 32 | 0.002558 | 0.002492 (0.000110) |
| 33 | 0.002270 | 0.002883 (0.000145) |
| 34 | 0.001918 | 0.002044 (0.000087) |
| 35 | 0.002640 | 0.002095 (0.000091) |
| 36 | 0.003576 | 0.003256 (0.000171) |
| 37 | 0.004401 | 0.003796 (0.000245) |
| 38 | 0.002728 | 0.003142 (0.000166) |
| 39 | 0.002187 | 0.002330 (0.000089) |
| 40 | 0.001561 | 0.002218 (0.000093) |

Table A.19 Across-item bias, SE and RMSE of the item difficulty parameter ($b$) average estimates based on 100 replications for 40 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | -2.0000 | -1.9868 | 0.1016 | 0.0132 | 0.1020 |
| 2 | -1.8974 | -1.9138 | 0.0954 | -0.0164 | 0.0963 |
| 3 | -1.7949 | -1.7863 | 0.0833 | 0.0086 | 0.0833 |
| 4 | -1.6923 | -1.6928 | 0.0933 | -0.0005 | 0.0928 |
| 5 | -1.5897 | -1.5924 | 0.0908 | -0.0026 | 0.0904 |
| 6 | -1.4872 | -1.4813 | 0.0654 | 0.0059 | 0.0653 |
| 7 | -1.3846 | -1.3802 | 0.0596 | 0.0045 | 0.0595 |
| 8 | -1.2821 | -1.2813 | 0.0724 | 0.0008 | 0.0720 |
| 9 | -1.1795 | -1.1869 | 0.0820 | -0.0074 | 0.0820 |
| 10 | -1.0769 | -1.0768 | 0.0742 | 0.0001 | 0.0738 |
| 11 | -0.9744 | -0.9756 | 0.0811 | -0.0012 | 0.0807 |
| 12 | -0.8718 | -0.8772 | 0.0751 | -0.0054 | 0.0749 |
| 13 | -0.7692 | -0.7669 | 0.0683 | 0.0023 | 0.0680 |
| 14 | -0.6667 | -0.6607 | 0.0656 | 0.0060 | 0.0656 |
| 15 | -0.5641 | -0.5669 | 0.0611 | -0.0028 | 0.0609 |
| 16 | -0.4615 | -0.4619 | 0.0436 | -0.0003 | 0.0434 |
| 17 | -0.3590 | -0.3549 | 0.0385 | 0.0041 | 0.0385 |
| 18 | -0.2564 | -0.2524 | 0.0375 | 0.0041 | 0.0376 |
| 19 | -0.1538 | -0.1581 | 0.0452 | -0.0042 | 0.0452 |
| 20 | -0.0513 | -0.0462 | 0.0562 | 0.0051 | 0.0562 |
| 21 | 0.0513 | 0.0533 | 0.0715 | 0.0021 | 0.0711 |
| 22 | 0.1538 | 0.1479 | 0.0476 | -0.0059 | 0.0478 |
| 23 | 0.2564 | 0.2555 | 0.0431 | -0.0009 | 0.0429 |
| 24 | 0.3590 | 0.3654 | 0.0562 | 0.0064 | 0.0562 |
| 25 | 0.4615 | 0.4635 | 0.0457 | 0.0020 | 0.0455 |
| 26 | 0.5641 | 0.5708 | 0.0425 | 0.0067 | 0.0428 |
| 27 | 0.6667 | 0.6661 | 0.0391 | -0.0005 | 0.0389 |
| 28 | 0.7692 | 0.7660 | 0.0484 | -0.0033 | 0.0482 |
| 29 | 0.8718 | 0.8766 | 0.0498 | 0.0048 | 0.0498 |
| 30 | 0.9744 | 0.9657 | 0.0745 | -0.0086 | 0.0746 |
| 31 | 1.0769 | 1.0891 | 0.0981 | 0.0121 | 0.0984 |
| 32 | 1.1795 | 1.1813 | 0.0620 | 0.0018 | 0.0618 |
| 33 | 1.2821 | 1.2747 | 0.0576 | -0.0073 | 0.0578 |
| 34 | 1.3846 | 1.3860 | 0.0947 | 0.0014 | 0.0942 |
| 35 | 1.4872 | 1.4827 | 0.0934 | -0.0045 | 0.0931 |
| 36 | 1.5897 | 1.5941 | 0.0666 | 0.0044 | 0.0664 |
| 37 | 1.6923 | 1.6893 | 0.0580 | -0.0030 | 0.0578 |
| 38 | 1.7949 | 1.8104 | 0.0657 | 0.0155 | 0.0672 |
| 39 | 1.8974 | 1.8943 | 0.1132 | -0.0032 | 0.1127 |
| 40 | 2.0000 | 2.0108 | 0.1136 | 0.0108 | 0.1135 |

Table A.20 Sample variance of the item difficulty ($b$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_b$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_b$ (SD) |
|---|---|---|
| 1 | 0.010328 | 0.010459 (0.001804) |
| 2 | 0.009101 | 0.009469 (0.001473) |
| 3 | 0.006941 | 0.007825 (0.001165) |
| 4 | 0.008700 | 0.008140 (0.001260) |
| 5 | 0.008244 | 0.008044 (0.001308) |
| 6 | 0.004278 | 0.005478 (0.000658) |
| 7 | 0.003552 | 0.004345 (0.000468) |
| 8 | 0.005235 | 0.005802 (0.000846) |
| 9 | 0.006728 | 0.005591 (0.000796) |
| 10 | 0.005499 | 0.005787 (0.000919) |
| 11 | 0.006578 | 0.008942 (0.001380) |
| 12 | 0.005637 | 0.006282 (0.000899) |
| 13 | 0.004669 | 0.004980 (0.000689) |
| 14 | 0.004310 | 0.003770 (0.000358) |
| 15 | 0.003736 | 0.003276 (0.000291) |
| 16 | 0.001901 | 0.002190 (0.000123) |
| 17 | 0.001481 | 0.001568 (0.000058) |
| 18 | 0.001408 | 0.001700 (0.000071) |
| 19 | 0.002044 | 0.002060 (0.000128) |
| 20 | 0.003159 | 0.002758 (0.000191) |
| 21 | 0.005105 | 0.003968 (0.000357) |
| 22 | 0.002269 | 0.002316 (0.000128) |
| 23 | 0.001861 | 0.001802 (0.000081) |
| 24 | 0.003154 | 0.002519 (0.000196) |
| 25 | 0.002085 | 0.002330 (0.000144) |
| 26 | 0.001808 | 0.002097 (0.000127) |
| 27 | 0.001530 | 0.001825 (0.000093) |
| 28 | 0.002341 | 0.002144 (0.000151) |
| 29 | 0.002477 | 0.002823 (0.000221) |
| 30 | 0.005552 | 0.005314 (0.000706) |
| 31 | 0.009628 | 0.010531 (0.002219) |
| 32 | 0.003850 | 0.003690 (0.000391) |
| 33 | 0.003318 | 0.003163 (0.000273) |
| 34 | 0.008970 | 0.009308 (0.001549) |
| 35 | 0.008732 | 0.009598 (0.001772) |
| 36 | 0.004435 | 0.003848 (0.000407) |
| 37 | 0.003360 | 0.003672 (0.000322) |
| 38 | 0.004321 | 0.005225 (0.000519) |
| 39 | 0.012810 | 0.011372 (0.001998) |
| 40 | 0.012901 | 0.016212 (0.003169) |

Table A.21 Across-item bias, SE and RMSE of the item slowness parameter ( $r$ ) average estimates based on 100 replications for 40 items and 2000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 0.7055 | 0.7133 | 0.0578 | 0.0078 | 0.0580 |
| 2 | 0.6760 | 0.6761 | 0.0545 | 0.0001 | 0.0543 |
| 3 | 0.5795 | 0.5834 | 0.0499 | 0.0039 | 0.0498 |
| 4 | 0.5340 | 0.5312 | 0.0511 | -0.0028 | 0.0509 |
| 5 | 0.5019 | 0.5112 | 0.0489 | 0.0093 | 0.0495 |
| 6 | 0.8980 | 0.9063 | 0.0513 | 0.0083 | 0.0517 |
| 7 | 1.0000 | 1.0015 | 0.0491 | 0.0015 | 0.0489 |
| 8 | 0.9230 | 0.9298 | 0.0476 | 0.0068 | 0.0478 |
| 9 | 0.8145 | 0.8153 | 0.0449 | 0.0008 | 0.0447 |
| 10 | 0.7860 | 0.7899 | 0.0413 | 0.0039 | 0.0413 |
| 11 | 0.5445 | 0.5483 | 0.0388 | 0.0038 | 0.0388 |
| 12 | 0.6780 | 0.6726 | 0.0358 | -0.0054 | 0.0360 |
| 13 | 0.8626 | 0.8681 | 0.0383 | 0.0055 | 0.0385 |
| 14 | 0.7920 | 0.7979 | 0.0384 | 0.0059 | 0.0386 |
| 15 | 0.7350 | 0.7383 | 0.0375 | 0.0033 | 0.0374 |
| 16 | 0.8120 | 0.8079 | 0.0399 | -0.0041 | 0.0399 |
| 17 | 0.8714 | 0.8734 | 0.0434 | 0.0020 | 0.0432 |
| 18 | 0.9230 | 0.9280 | 0.0425 | 0.0050 | 0.0426 |
| 19 | 0.9496 | 0.9434 | 0.0435 | -0.0062 | 0.0437 |
| 20 | 0.6780 | 0.6835 | 0.0357 | 0.0055 | 0.0360 |
| 21 | 0.5249 | 0.5281 | 0.0396 | 0.0032 | 0.0396 |
| 22 | 0.5950 | 0.5923 | 0.0374 | -0.0027 | 0.0373 |
| 23 | 0.5350 | 0.5354 | 0.0382 | 0.0004 | 0.0380 |
| 24 | 0.6230 | 0.6277 | 0.0405 | 0.0047 | 0.0405 |
| 25 | 0.6874 | 0.6897 | 0.0428 | 0.0023 | 0.0427 |
| 26 | 0.6920 | 0.6956 | 0.0400 | 0.0036 | 0.0399 |
| 27 | 0.6227 | 0.6174 | 0.0367 | -0.0053 | 0.0369 |
| 28 | 0.7970 | 0.7946 | 0.0443 | -0.0024 | 0.0441 |
| 29 | 0.8326 | 0.8389 | 0.0426 | 0.0063 | 0.0429 |
| 30 | 0.8010 | 0.7988 | 0.0434 | -0.0022 | 0.0433 |
| 31 | 0.8298 | 0.8274 | 0.0369 | -0.0024 | 0.0367 |
| 32 | 0.7230 | 0.7306 | 0.0457 | 0.0076 | 0.0461 |
| 33 | 0.7892 | 0.7882 | 0.0487 | -0.0010 | 0.0484 |
| 34 | 0.8930 | 0.8938 | 0.0382 | 0.0008 | 0.0380 |
| 35 | 0.9110 | 0.9050 | 0.0425 | -0.0060 | 0.0428 |
| 36 | 0.7520 | 0.7498 | 0.0559 | -0.0022 | 0.0557 |
| 37 | 0.6951 | 0.6858 | 0.0634 | -0.0093 | 0.0638 |
| 38 | 0.5930 | 0.6103 | 0.0592 | 0.0173 | 0.0614 |
| 39 | 0.5439 | 0.5435 | 0.0591 | -0.0004 | 0.0588 |
| 40 | 0.6250 | 0.6186 | 0.0479 | -0.0064 | 0.0481 |

Table A.22 Sample variance of the item slowness ($r$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}_r$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 2000 examinees

| Item | $S^2$ | $\overline{I}_r$ (SD) |
|---|---|---|
| 1 | 0.003341 | 0.003002 (0.000165) |
| 2 | 0.002973 | 0.002932 (0.000157) |
| 3 | 0.002493 | 0.002838 (0.000145) |
| 4 | 0.002607 | 0.002494 (0.000144) |
| 5 | 0.002392 | 0.002264 (0.000109) |
| 6 | 0.002629 | 0.002452 (0.000116) |
| 7 | 0.002412 | 0.002487 (0.000114) |
| 8 | 0.002265 | 0.001967 (0.000096) |
| 9 | 0.002013 | 0.001848 (0.000076) |
| 10 | 0.001709 | 0.001694 (0.000069) |
| 11 | 0.001507 | 0.001474 (0.000055) |
| 12 | 0.001282 | 0.001500 (0.000051) |
| 13 | 0.001465 | 0.001491 (0.000050) |
| 14 | 0.001473 | 0.001496 (0.000050) |
| 15 | 0.001405 | 0.001475 (0.000045) |
| 16 | 0.001594 | 0.001523 (0.000048) |
| 17 | 0.001880 | 0.001578 (0.000048) |
| 18 | 0.001805 | 0.001489 (0.000044) |
| 19 | 0.001893 | 0.001411 (0.000039) |
| 20 | 0.001275 | 0.001353 (0.000037) |
| 21 | 0.001570 | 0.001322 (0.000035) |
| 22 | 0.001400 | 0.001384 (0.000041) |
| 23 | 0.001460 | 0.001462 (0.000045) |
| 24 | 0.001637 | 0.001431 (0.000049) |
| 25 | 0.001835 | 0.001498 (0.000053) |
| 26 | 0.001598 | 0.001617 (0.000063) |
| 27 | 0.001347 | 0.001804 (0.000075) |
| 28 | 0.001962 | 0.001845 (0.000083) |
| 29 | 0.001815 | 0.001827 (0.000083) |
| 30 | 0.001886 | 0.001619 (0.000072) |
| 31 | 0.001358 | 0.001502 (0.000059) |
| 32 | 0.002091 | 0.002165 (0.000116) |
| 33 | 0.002370 | 0.002610 (0.000144) |
| 34 | 0.001458 | 0.001807 (0.000085) |
| 35 | 0.001810 | 0.001920 (0.000100) |
| 36 | 0.003127 | 0.003602 (0.000238) |
| 37 | 0.004020 | 0.004411 (0.000280) |
| 38 | 0.003506 | 0.003978 (0.000264) |
| 39 | 0.003495 | 0.002579 (0.000180) |
| 40 | 0.002291 | 0.002385 (0.000140) |

Table A.23 Across-item bias, SE and RMSE of the parameter $(\eta, g, \sigma, \sigma_s)$ average estimates based on 100 replications for 40 items and 2000 examinees

| Parameter | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| $\eta$ | 0.002 | 0.0020 | 0.0005 | -0.00001 | 0.0005 |
| $g$ | 0.5 | 0.4990 | 0.0145 | -0.0010 | 0.0144 |
| $\sigma$ | 1 | 0.9998 | 0.0027 | -0.0002 | 0.0027 |
| $\sigma_s$ | 1 | 0.9997 | 0.0123 | -0.0003 | 0.0123 |

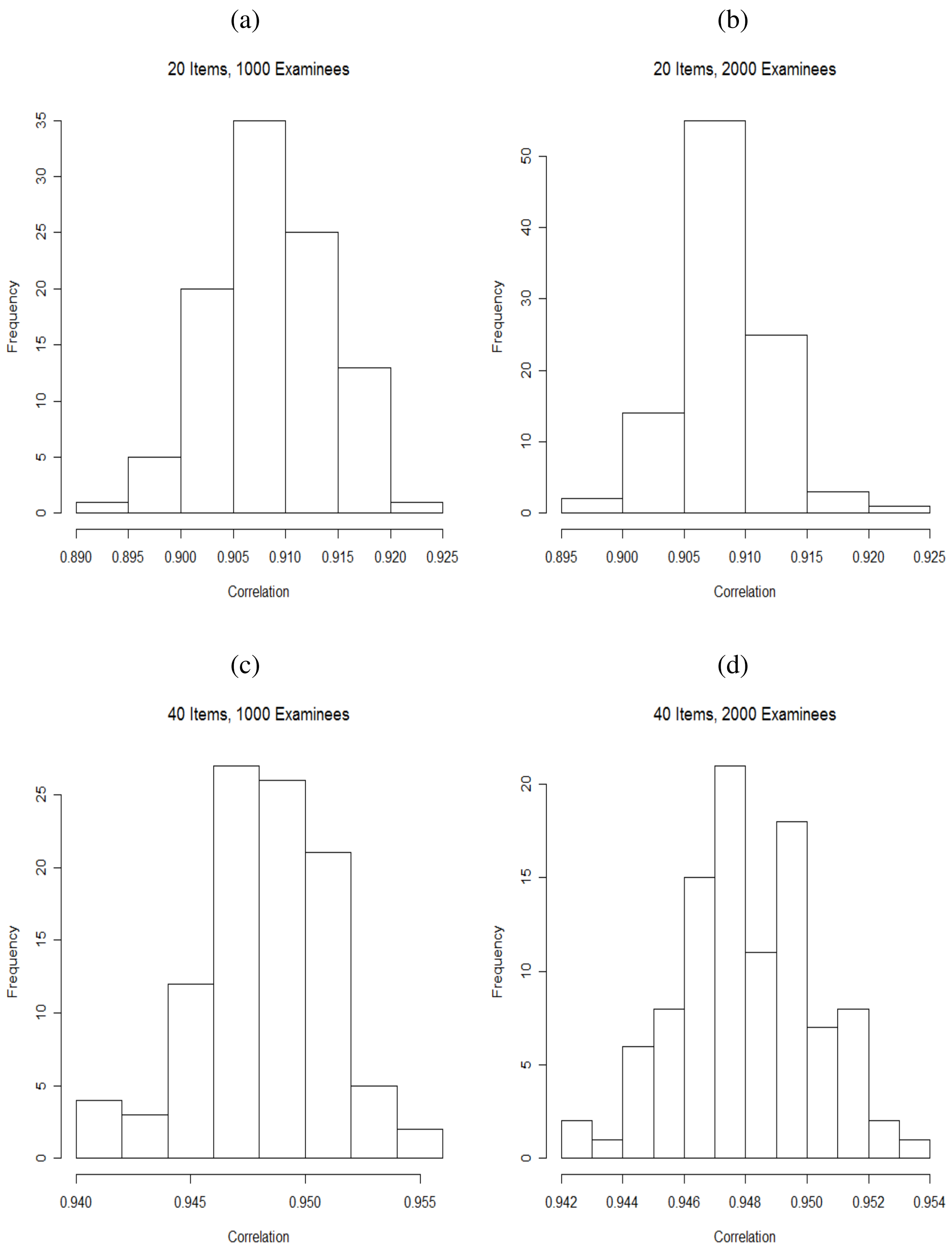
Table A.24 Sample variance of the parameter ($\eta, g, \sigma, \sigma_s$) estimates ($S^2$) and mean of the variance estimates based on the Fisher information ($\overline{I}$) and corresponding standard deviation (SD) over 100 simulation replications for 40 items and 2000 examinees

| Parameter | $S^2$ | $\overline{I}$ (SD) |
|---|---|---|
| $\eta$ | 0.00000029 | 0.00000028 (0.00000004) |
| $g$ | 0.00020889 | 0.00026227 (0.00001927) |
| $\sigma$ | 0.00000724 | 0.00000643 (0.00000003) |
| $\sigma_s$ | 0.00015216 | 0.00014445 (0.00000676) |

# APPENDIX B

## DETAILED ANALYSIS OF SIMULATION DESIGN 2 – PERSON PARAMETER ESTIMATION USING MAP

The histograms of the distribution of correlations between true and estimated person parameters for all four simulation conditions are presented. Figure B.1 displays the histograms of correlations between true and estimated ability parameters $\left(\theta, \hat{\theta}\right)$ from 100 replications across four simulation conditions, (a) 20 items for 1000 examinees, (b) 20 items for 2000 examinees, (c) 40 items for 1000 examinees, and (d) 40 items for 2000 examinees. Figure B.2 exhibits the histograms of correlations of between true and estimated person slowness parameters $\left(s, \hat{s}\right)$ from 100 replications across four simulation conditions, (a) 20 items for 1000 examinees, (b) 20 items for 2000 examinees, (c) 40 items for 1000 examinees, and (d) 40 items for 2000 examinees.


(a)

(b)

(c)

(d)

Figure B.1 Histograms of correlations between true and estimated ability parameters $(\theta, \hat{\theta})$ from 100 replications across four simulation conditions, (a) 20 items for 1000 examinees, (b) 20 items for 2000 examinees, (c) 40 items for 1000 examinees, and (d) 40 items for 2000 examinees

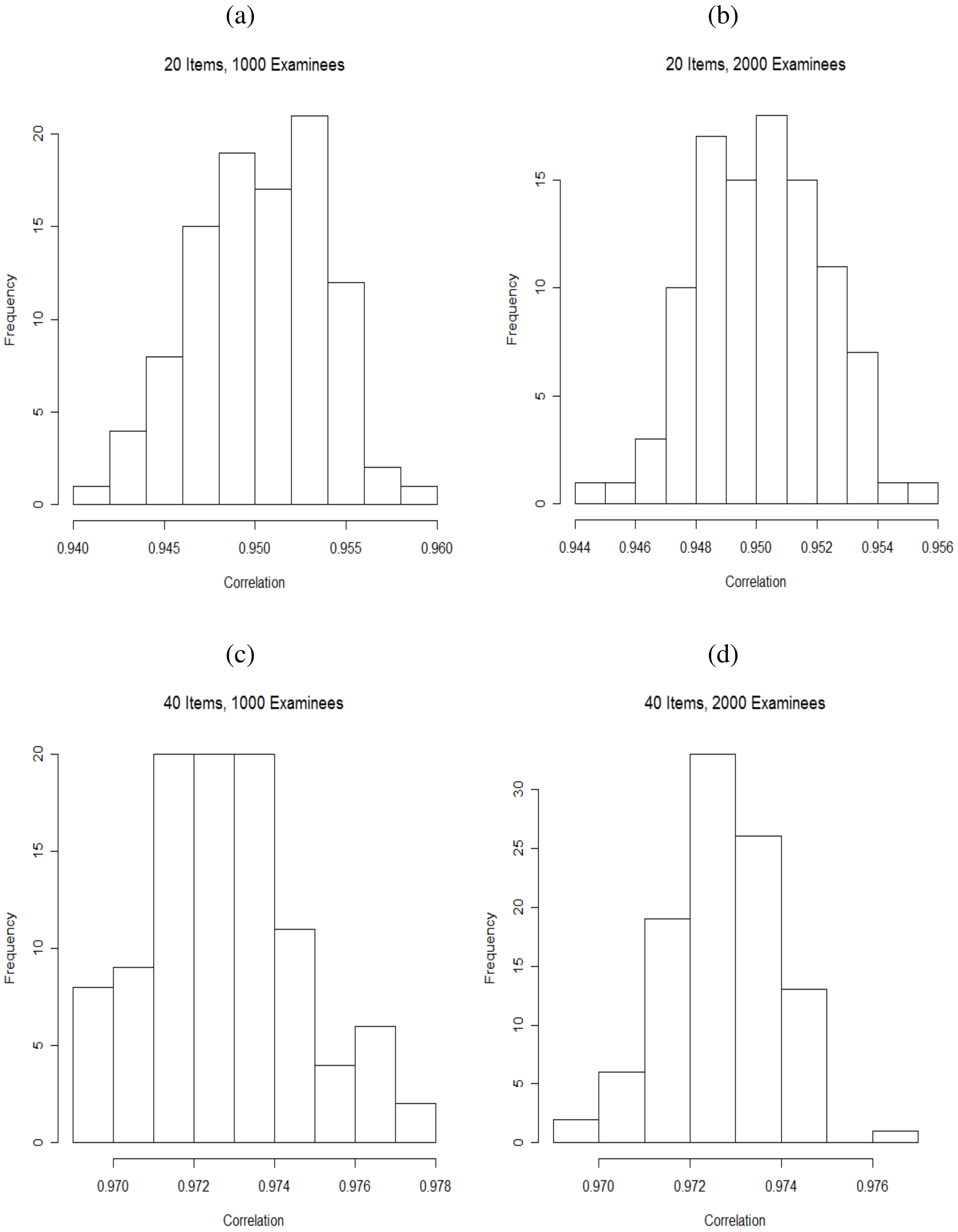


Figure B.2 Histograms of correlations between true and estimated person slowness parameters $\left(s, \hat{s}\right)$ from 100 replications across four simulation conditions, (a) 20 items for 1000 examinees, (b) 20 items for 2000 examinees, (c) 40 items for 1000 examinees, and (d) 40 items for 2000 examinees

## APPENDIX C

## DETAILED ANALYSIS OF SIMULATION DESIGN 3 – PERSON PARAMETER RECOVERY USING MML AND MAP

In simulation design 3 – person parameter recovery, two scenarios were investigated to see how well ability and person slowness parameters are estimated using MAP procedure. Scenario 1 is when the true item parameters are known. Scenario 2 is when item parameters are estimated. A simulation for 20 items and 1000 examinees is performed. Here, Tables C.1, C.2 and C.3 contain the item parameter estimation results from scenario 2. In addition, from scenario 2, additional parameter estimation results are given in Table C.4.

Tables C.1, C.2 and C.3 contain bias, SE and RMSE and average estimates of item discrimination parameter ($a$), item difficulty parameter ($b$) and item slowness parameter ($r$) based on 100 replications for 20 items and 1000 examinees. Table C.4 include bias, SE and RMSE of the parameter ($\eta, g, \sigma, \sigma_s$) average estimates based on 100 replications for 20 items and 1000 examinees.

The histograms of the distribution of correlations between true and estimated person parameters for all four simulation conditions are presented. In Figure C.1, histograms of correlations between true and estimated ability $\left(\theta, \hat{\theta}\right)$ parameters from 100 replications across two simulation conditions for 20 items for 1000 examinees. (a) item parameters are known and (b) item parameters are estimated. In Figure C.2, histograms of correlations between true and estimated person slowness $\left(s, \hat{s}\right)$ parameters from 100 replications across two simulation conditions for 20 items for 1000 examinees. (a) item parameters are known and (b) item parameters are estimated.

Table C.1 Across-item bias, SE and RMSE of the item discrimination parameter ($a$) average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 1.0036 | 1.0018 | 0.0843 | -0.0018 | 0.0838 |
| 2 | 1.0650 | 1.0553 | 0.0760 | -0.0097 | 0.0762 |
| 3 | 0.9497 | 0.9566 | 0.0660 | 0.0069 | 0.0660 |
| 4 | 1.2234 | 1.2164 | 0.0672 | -0.0070 | 0.0672 |
| 5 | 0.9407 | 0.9370 | 0.0816 | -0.0037 | 0.0813 |
| 6 | 0.6576 | 0.6543 | 0.0609 | -0.0033 | 0.0607 |
| 7 | 0.8279 | 0.8310 | 0.0695 | 0.0031 | 0.0692 |
| 8 | 0.9844 | 0.9921 | 0.0708 | 0.0077 | 0.0709 |
| 9 | 1.5805 | 1.5802 | 0.0834 | -0.0003 | 0.0830 |
| 10 | 1.2155 | 1.2162 | 0.0747 | 0.0007 | 0.0743 |
| 11 | 0.7860 | 0.7744 | 0.0649 | -0.0116 | 0.0656 |
| 12 | 1.3716 | 1.3727 | 0.0817 | 0.0011 | 0.0813 |
| 13 | 1.2027 | 1.2012 | 0.0775 | -0.0015 | 0.0771 |
| 14 | 1.5569 | 1.5489 | 0.0801 | -0.0080 | 0.0801 |
| 15 | 1.2431 | 1.2460 | 0.0800 | 0.0029 | 0.0797 |
| 16 | 0.6453 | 0.6321 | 0.0609 | -0.0132 | 0.0620 |
| 17 | 1.4352 | 1.4486 | 0.0819 | 0.0134 | 0.0826 |
| 18 | 0.8156 | 0.8168 | 0.0730 | 0.0012 | 0.0726 |
| 19 | 1.7798 | 1.7796 | 0.1020 | -0.0002 | 0.1014 |
| 20 | 0.9110 | 0.9014 | 0.0737 | -0.0096 | 0.0740 |

Table C.2 Across-item bias, SE and RMSE of the item difficulty parameter ($b$) average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | -2.0000 | -2.0085 | 0.1632 | -0.0085 | 0.1626 |
| 2 | -1.7895 | -1.8008 | 0.1155 | -0.0113 | 0.1154 |
| 3 | -1.5789 | -1.5749 | 0.1247 | 0.0041 | 0.1241 |
| 4 | -1.3684 | -1.3895 | 0.1006 | -0.0211 | 0.1023 |
| 5 | -1.1579 | -1.1815 | 0.1197 | -0.0236 | 0.1215 |
| 6 | -0.9474 | -0.9647 | 0.1404 | -0.0173 | 0.1407 |
| 7 | -0.7368 | -0.7399 | 0.1011 | -0.0030 | 0.1007 |
| 8 | -0.5263 | -0.5192 | 0.0764 | 0.0071 | 0.0764 |
| 9 | -0.3158 | -0.3139 | 0.0548 | 0.0019 | 0.0546 |
| 10 | -0.1053 | -0.0955 | 0.0589 | 0.0097 | 0.0594 |
| 11 | 0.1053 | 0.1148 | 0.0832 | 0.0095 | 0.0833 |
| 12 | 0.3158 | 0.3285 | 0.0537 | 0.0127 | 0.0549 |
| 13 | 0.5263 | 0.5328 | 0.0709 | 0.0064 | 0.0708 |
| 14 | 0.7368 | 0.7400 | 0.0573 | 0.0032 | 0.0571 |
| 15 | 0.9474 | 0.9412 | 0.0797 | -0.0062 | 0.0795 |
| 16 | 1.1579 | 1.1848 | 0.1665 | 0.0269 | 0.1679 |
| 17 | 1.3684 | 1.3661 | 0.0826 | -0.0023 | 0.0822 |
| 18 | 1.5789 | 1.5900 | 0.1517 | 0.0111 | 0.1513 |
| 19 | 1.7895 | 1.8098 | 0.0898 | 0.0203 | 0.0917 |
| 20 | 2.0000 | 2.0243 | 0.1803 | 0.0243 | 0.1810 |

Table C.3 Across-item bias, SE and RMSE of the item slowness parameter ( $r$ ) average estimates based on 100 replications for 20 items and 1000 examinees

| Item | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| 1 | 0.7055 | 0.7248 | 0.0887 | 0.0193 | 0.0904 |
| 2 | 0.5795 | 0.6046 | 0.0815 | 0.0251 | 0.0849 |
| 3 | 0.5019 | 0.5206 | 0.0773 | 0.0187 | 0.0791 |
| 4 | 1.0000 | 1.0146 | 0.0764 | 0.0146 | 0.0774 |
| 5 | 0.8145 | 0.8157 | 0.0660 | 0.0012 | 0.0657 |
| 6 | 0.5445 | 0.5568 | 0.0554 | 0.0123 | 0.0565 |
| 7 | 0.8626 | 0.8753 | 0.0583 | 0.0127 | 0.0594 |
| 8 | 0.7350 | 0.7494 | 0.0574 | 0.0144 | 0.0589 |
| 9 | 0.8714 | 0.8836 | 0.0610 | 0.0122 | 0.0619 |
| 10 | 0.9496 | 0.9661 | 0.0598 | 0.0165 | 0.0617 |
| 11 | 0.5249 | 0.5320 | 0.0497 | 0.0071 | 0.0500 |
| 12 | 0.5350 | 0.5498 | 0.0576 | 0.0148 | 0.0592 |
| 13 | 0.6874 | 0.6903 | 0.0582 | 0.0029 | 0.0580 |
| 14 | 0.6227 | 0.6210 | 0.0653 | -0.0017 | 0.0650 |
| 15 | 0.8326 | 0.8286 | 0.0599 | -0.0040 | 0.0598 |
| 16 | 0.8298 | 0.8292 | 0.0518 | -0.0006 | 0.0515 |
| 17 | 0.7892 | 0.7887 | 0.0741 | -0.0005 | 0.0737 |
| 18 | 0.9110 | 0.9130 | 0.0681 | 0.0020 | 0.0678 |
| 19 | 0.6951 | 0.6951 | 0.1155 | 0.00001 | 0.1149 |
| 20 | 0.5439 | 0.5377 | 0.0855 | -0.0062 | 0.0853 |

Table C.4 Across-item bias, SE and RMSE of the parameter $(\eta, g, \sigma, \sigma_s)$ average estimates based on 100 replications for 20 items and 1000 examinees

| Parameter | True | Estimates | SE | Bias | RMSE |
|---|---|---|---|---|---|
| $\eta$ | 0.002 | 0.0019 | 0.0010 | -0.00007 | 0.0010 |
| $g$ | 0.5 | 0.4934 | 0.0299 | -0.00657 | 0.0305 |
| $\sigma$ | 1 | 0.9991 | 0.0053 | -0.00088 | 0.0054 |
| $\sigma_s$ | 1 | 0.9988 | 0.0183 | -0.00116 | 0.0183 |

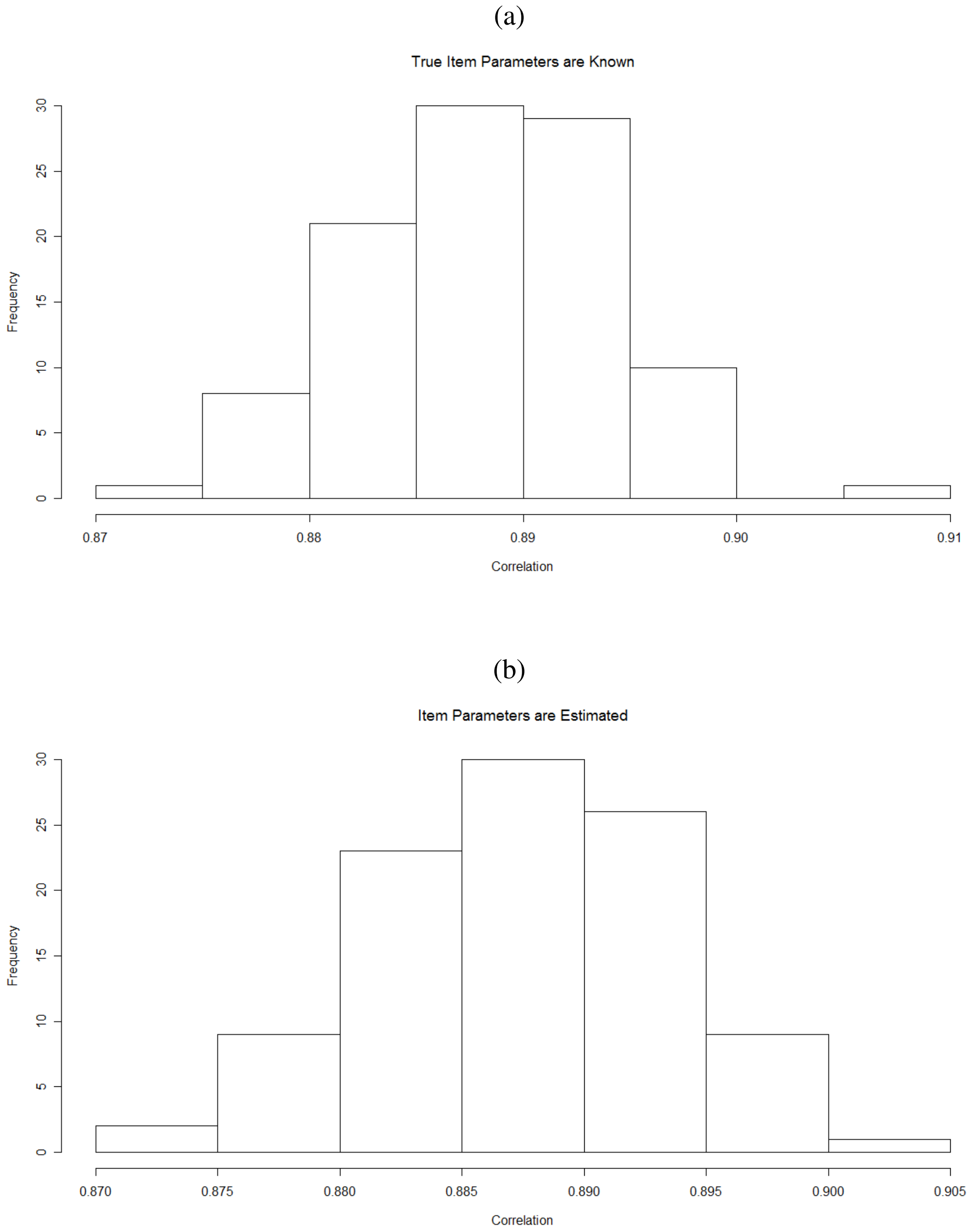
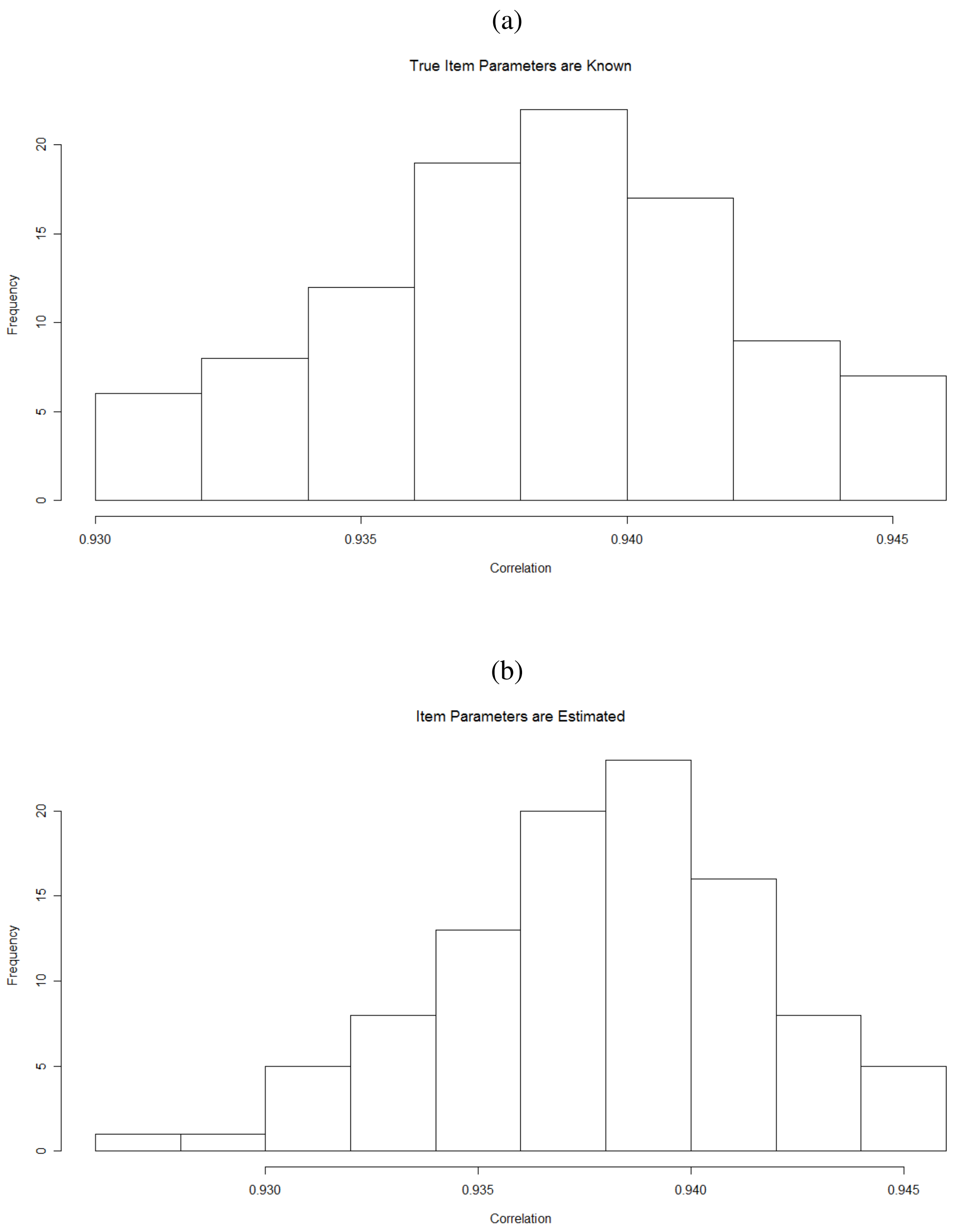


Figure C.1 Histograms of correlations between true and estimated ability $(\theta, \hat{\theta})$ parameters from 100 replications across two simulation conditions for 20 items for 1000 examinees. (a) item parameters are known and (b) item parameters are estimated



(a)

True Item Parameters are Known

(b)

Item Parameters are Estimated

Figure C.2 Histograms of correlations between true and estimated person slowness $(s, \hat{s})$ parameters from 100 replications across two simulation conditions for 20 items for 1000 examinees. (a) item parameters are known and (b) item parameters are estimated

# APPENDIX D

## ADDITIONAL ANALYSIS ON PERSON PARAMETERS

An additional analysis on person parameters, i.e. ability $(\theta)$ and person slowness $(s)$, is included in Appendix D. Table D.1 shows the correlation between true ability and true person slowness parameters for four simulation conditions, 20/ 40 items for 1000/ 2000 examinees. In the Table D.1, the correlation values are close to zero. It means that the simulated true ability and true person slowness data are not correlated for four simulation conditions. Thus, it corresponds with the independence assumption between ability and person slowness parameters.

Table D.1 Correlation between true ability and true person slowness parameters ($\theta, s$)

| Item | Examinee | Correlation<br>$\theta, s$ |
|---|---|---|
| 20 | 1000 | 0.0380 |
|  | 2000 | 0.0314 |
| 40 | 1000 | -0.0406 |
|  | 2000 | -0.0046 |

# REFERENCES

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/ Cole Publishing Company.

Bock, D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika*, 46, 443-459.

Bridgeman, B., Cline, F., & Hessinger, J. (2003). *Effect of extra time on GRE® quantitative and verbal scores*. (ETS Report RR-03-13, GRE-00-03P). Princeton, NJ: Educational Testing Service.

Bridgeman, B., McBride, A., & Monaghan, W. (2004). *R&D connections - testing and time limits*. (ETS Report RDC-01). Princeton, NJ: Educational Testing Service.

Broyden, C. G., (1970). The Convergence of a class of double-rank minimization algorithms, *Journal of the Institute of Mathematics and Its Applications*, 6, 76-90.

Casella, G., & Berger, R.L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Embretson, S.E. (1985). Introduction to the problem of test design. In S.E. Embretson (Ed.) *Test design: Developments in psychology and psychometrics*. (pp. 3-17). Orlando, FL: Academic press.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologist*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Fletcher, R., (1970). A new approach to variable metric algorithms, *Computer Journal*, 13, 317-322.

Folk, V.G., & Smith, R.L. (2002). Models for delivery of CBTs, In C.N. Mills, M.T. Potenza, J.J. Fremer & W.C. Ward (Eds.) *Computer-based testing: Building the foundation for future assessment.*(pp. 41-66). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Furneaux, W.D. (1961). Intellectual abilities and problem solving behavior. In H.J. Eysenck (Ed.), *The handbook of abnormal psychology*, (pp. 167-192). London, England: Pitman.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means, *Mathematics of Computation*, 24, 23-26.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory.* Boston, MA: Kluwer Nijhoff publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE publications.

Hopkins, K.D. (1998). *Educational and psychological measurement and evaluation*. Needham Heights, MA: Allyn & Bacon.

Ingrisone, J.N. (2008). *Modeling the joint distribution of response accuracy and response time*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hilsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lord, F.M. & Novick, M.R. (1980). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R.D. (1986). *Response times*. New York, NY: Oxford University Press.

Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200-219.

Parshall, C.G., Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer-Verlag.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org.

Roskam, E.E.(1997). Models for speed and time-limit tests. In W.J. Van der Linden & R.K. Hambleton (Eds.) *Handbook of modern item response theory*. (pp.187-208). New York, NY: Springer-Verlag.

Samejima, F. (1973). Homogeneous case of the continuous response level. *Psychometrika*, 38, 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111-121.

Samejima, F. (1983). *A general model for the homogeneous case of the continuous response* (ONR Research Report 83-3). Arlington, VA: Office of Naval Research, Personnel and Training Research Program.

Schleiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, 19, 18-38.

Schleiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S.E. Embretson (Ed.) *Test design: Developments in psychology and psychometrics*. (pp. 219-244). Orlando, FL: Academic press.

Schnipke, D.L., & Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement*, 34, 213-232.

Schnipke, D.L., & Scrams, D.J. (1999). *Representing response time information in item banks*. (LSAC Computerized Testing Report 97-09). Newtown, PA: Law School Admission Council.

Schnipke, D.L., & Scrams, D.J. (2002). Exploring issues of examinee behavior: insights gained from response-time analyses. In C.N. Mills, M.T. Potenza, J.J. Fremer & W.C. Ward (Eds.) *Computer-based testing: Building the foundation for future assessment.* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Scrams, D.J., & Schnipke, D.L. (1997, March). *Making use of item response times in standardized tests: are accuracy and speed measuring the same thing?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization, *Mathematics of Computation*, 24, 647-656.

Storms, G., & Delbeke, L. (1992). The irrelevance of distributional assumptions on reaction times in multidimensional scaling of same/ different judgment tasks. *Psychometrika*, 57, 599-614.

Tatsuoka, K.K., & Tatsuoka, M.M. (1980). A model for incorporating response-time data in scoring achievement tests. In D.Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236-256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D. (1983). Timed testing: an approach using item response theory. In D.J. Weiss (Ed.). *New horizons in testing*. (pp.179-203). New York, NY: Academic Press.

Titterington, D.M., Smith, A.F.M., & Markov, U.E. (1985). *Statistical analysis of finite mixture distributions*. New York, NY: John Wiley & Sons.

van Bruekelen, G.J.P. (1997). Separability of item and person parameters in response time models. *Psychometrika*, 62, 525-544.

van der Linden, W. (2006). Normal models for response times on test items. (LSAC Computerized Testing Report 04-08). Princeton, NJ: Law School Admission Council.

van der Linden,W.J., & Hambleton, R.K. (1997). Models for response time or multiple attempts on items: introduction. In W.J. van der Linden, & R.K. Hambleton (Eds.) *Handbook of modern item response theory*. (pp.165-168). New York, NY: Springer- Verlag.

van der Linden, W., Scrams, D.J., & Schnipke, D.L. (1999). Using response time constraints for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23, 195-210.

van der Linden, W., & Krimpen-Stoop, E.M.L.A. (2003). Using response times to detect abberant responses in computerized adaptive testing. *Psychometrika*, 68, 251-265.

van der Linden, W., & Krimpen-Stoop, E.M.L.A. (2005). *Using response times to detect abberant responses in computerized adaptive testing*. (LSAC Computerized Testing Report 01-02). Princeton, NJ: Law School Admission Council.

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G.H. (1997). A logistic model for time-limit tests. In W.J. van der Linden, & R.K. Hambleton (Eds.) *Handbook of modern item response theory*. (pp.169-186). New York, NY: Springer-Verlag.

Wang, T. (2006). *A model for the joint distribution of item response and response time using one-parameter Weibull distribution* (CASMA Research Report 20). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Wang, T., & Hanson, B.A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29, 323- 339.

Wang, T., & Zhang, J. (2006). Optimal partitioning of testing time: theoretical properties and practical implications. *Psychometrika*, 71, 105-120.

Weiss, D.J., & Yoes, M.E. (1990). Item response theory. In R.K. Hambleton, & J.N. Zaal. *Advances in educational and psychological testing*. Boston, MA: Kluwer Academic Publishers.

White, P.O. (1973). Individual difference in speed, accuracy and persistence. In H.J. Eysenck (Ed.), *The measurement of intelligence*. Lancaster, England: Medical and Technical Publishing Co.

White, P.O. (1979, June). *A latent trait model for individual differences in speed, accuracy, and persistence*. Paper presented at the annual meeting of the Psychometric Society, Monterey, California.

Woodruff, D. & Hanson, B.A. (1997, June). *Estimation for item response models using the EM algorithm for finite mixtures*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, Tennessee.

Yen, W.M., & Fitzpatrick, A.R. (2006). Item response theory. In R.L. Brennan (Ed.). *Educational measurement*, 4th Ed. Westport, CT: American Council on Education and Praeger Publishers.

# BIOGRAPHICAL SKETCH

Soo Jeong Ingrisone was born and raised in Seoul, Korea. For her undergraduate degree, she studied art history at Hong-Ik University in Seoul, Korea. During her undergraduate education, she was introduced to the philosophy of beauty, which led her to pursue a master's degree in Philosophy. At Ludwig-Maximilians-University in Munich, Germany, she majored in philosophy and minored in sociology and art history. By the end of her education she became a Kantian. In the course of her Ph.D. work at Florida State University, she was introduced to various research methods which she found very exciting and intriguing. Consequently, it led her to the study of measurement and statistics. During her studies, she became fascinated by the elegance of mathematics and the logic of the computer programming, and developed a passion for solving applied statistical problems in educational measurement. She specializes in the area of modeling the joint distribution of response accuracy and response time as well as investigating procedures to estimate ability and item parameters.

She is fluent in three different languages, Korean, German and English, and has basic knowledge in Latin, Japanese, Chinese and French. She enjoys good food, literatures, movies, classical music, impressionism art, playing piano, painting and travel. She is happily married with one child.